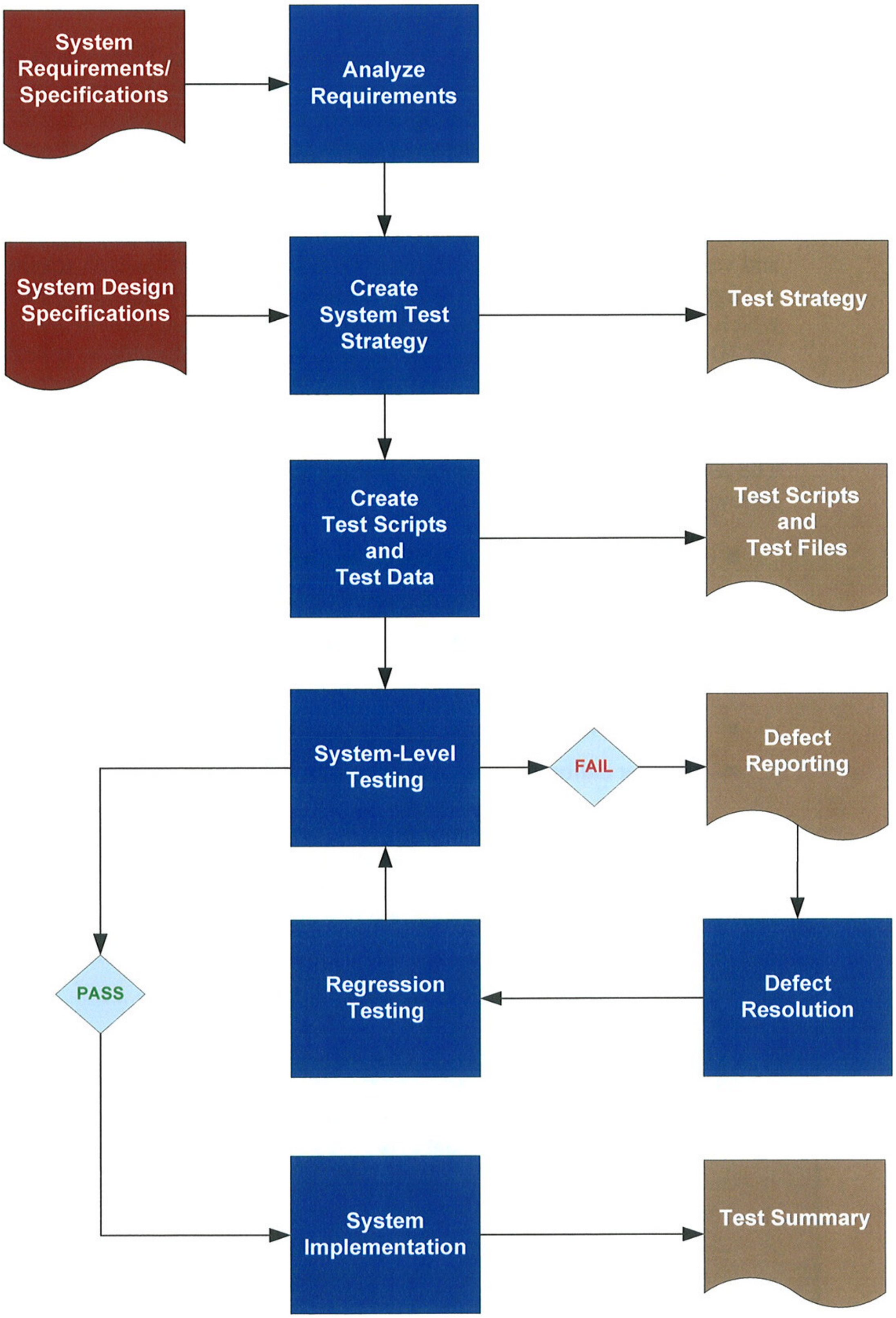

System Requirements/ Specifications
Analyze Requirements
System Design Specifications
Create System Test Strategy
Test Strategy
Create Test Scripts and Test Data
Test Scripts and Test Files
System-Level Testing
FAIL
Defect Reporting
PASS
Regression Testing
Defect Resolution
System Implementation
Test Summary

Figure 4–52. Quality Assurance System Testing Model

Test Decks

Prior to any NeSA test materials returning to DRC, the Software Quality Assurance staff will perform extensive tests to ensure all scanned data (including demographic and multiple-choice responses) are captured and accurately stored in a secure database environment. Each record in the database will be independently verified against the test decks for validation.

The analysts will follow a software testing methodology that thoroughly evaluates and verifies the scanning and scoring system and verifies that each scanner is configured and set up for the NeSA. This process includes validating test decks, which will be comprised of answer sheets with and without student and school pre-ID information for each form of the test. The test decks will be specifically gridded to include a variety of possible student response permutations and combinations.

The test decks will be processed completely through DRC's systems to verify the following:

- Readability of security, student, and school barcodes.
- Data capture of pre-ID and barcode information.
- Accurate capture of district and school codes.
- Consistent data capture on all scanners.
- Accurate scan positions on all documents and forms.
- Scanner calibration and hardware functionality.

The Software Quality Assurance staff will also perform a validation of all production data processed through the system. **Each student record will be verified for accuracy** to ensure high quality data file development and reporting. DRC's scanning quality procedures are presented in Figure 4–53.

- **Test Decks—**DRC will process test decks configured for the NeSA through the production systems.
- **Calibration—**Daily calibration and scanner cleaning processes will be conducted to ensure read level consistency.
- **Standard edit processes—**Every scannable document will be processed through edit programs to detect potential errors (double marks, smudge marks, omits, etc.)
- **Multiple reviews—**The Document Processing Supervisor will conduct a review of the entire first batch prior to full production to ensure error-free processing.
- **Quality control reports—**Daily quality control reports for each editor will be reviewed by the Document Processing Supervisor to monitor the accuracy of the online editing process.

Figure 4–53. Scanning Quality Procedures

iii. Scanning Database

Creation of Data Files

Student responses to multiple-choice items, as well as demographic information, will be captured as images and preserved for use during the image scoring process. Information embedded in the student precode label or the district/school label will also be captured during scanning. This information will link back to the NDE Student ID record or to the site at which the student tested (if a school/district label was used). Booklet counts and page integrity will be maintained throughout the scanning process by storing data in a Relational Database Management System (RDMS) using unique identifiers that link each image to a single, individual record, preserving school/district and other identification and demographic information. A relational database significantly increases system flexibility and provides for robust data analysis capabilities.

After the demographic information and multiple-choice data pass all pre-defined editing processes, the data and information is available for student-level processing and scoring (please see *Subheading 6, Scoring,* for more information on DRC's scoring processes).

Data File Accuracy

Below, we present an overview of the processes and methods DRC will use to ensure the accuracy and completeness of student data and associated data files:

- All student answer sheets returned to DRC will be scored. Multiple-choice items and demographic information will be image scanned and the original scanned data will be converted into a master student file. All student information and score results are kept secure and confidential throughout the scanning, scoring, and reporting processes.
  - All answer sheets will be processed, scored and reported. No invalidation process was requested in the RFP or Questions and Answers. Should NDE desire this service, costs can be provided upon award.
- Record counts will be verified against the counts from the Document Processing staff to ensure all students are accounted for in the file. Additionally, a detailed review of the error-tracking log will be performed to ensure any discrepancies are resolved before proceeding with the scoring routines.
- The scoring process will include the scoring of multiple-choice items to the answer key and the aggregation of any raw scores from online testing. Using the raw scores, scaled scores will be calculated.
- After scanning and scoring, DRC's Software Quality Assurance will perform an item response frequency analysis on both initial and complete data sets. DRC Psychometrics staff will perform item response frequency analysis, independent foil analysis, and differential item functioning

analysis (please see *Subheading 7, Analyses,* for a thorough discussion of DRC's proposed data analysis procedures).

- Raw-to-scale score conversion tables and cut points based on pre-equating are provided to the DRC Information Systems Team. The conversion table allows each student's raw score to be converted into a scaled score. The cut points are used to assign each student to a proficiency category.
- Additional reporting software will contain the procedures for sorting and summarizing data. Our Software Quality Assurance staff will ensure the quality of school-, district-, and state-level data and make certain that each record is verified for completeness and accuracy. Quality checks will be performed on the data placement and data file formatting for each data element to be displayed on the reports. All data elements will be verified back to the production data file and the data processing rules. Senior Software Quality Assurance Analysts will conduct another review to ensure methodology, processes, and procedures are followed and verify that the data files are approved prior to report production.
- All data files for the NeSA reports will be quality checked for accuracy and completeness by DRC Quality Assurance Analysts.
- All data file design, development, and enhancement efforts will be done in close association with NDE to ensure that all requirements for reporting are met.

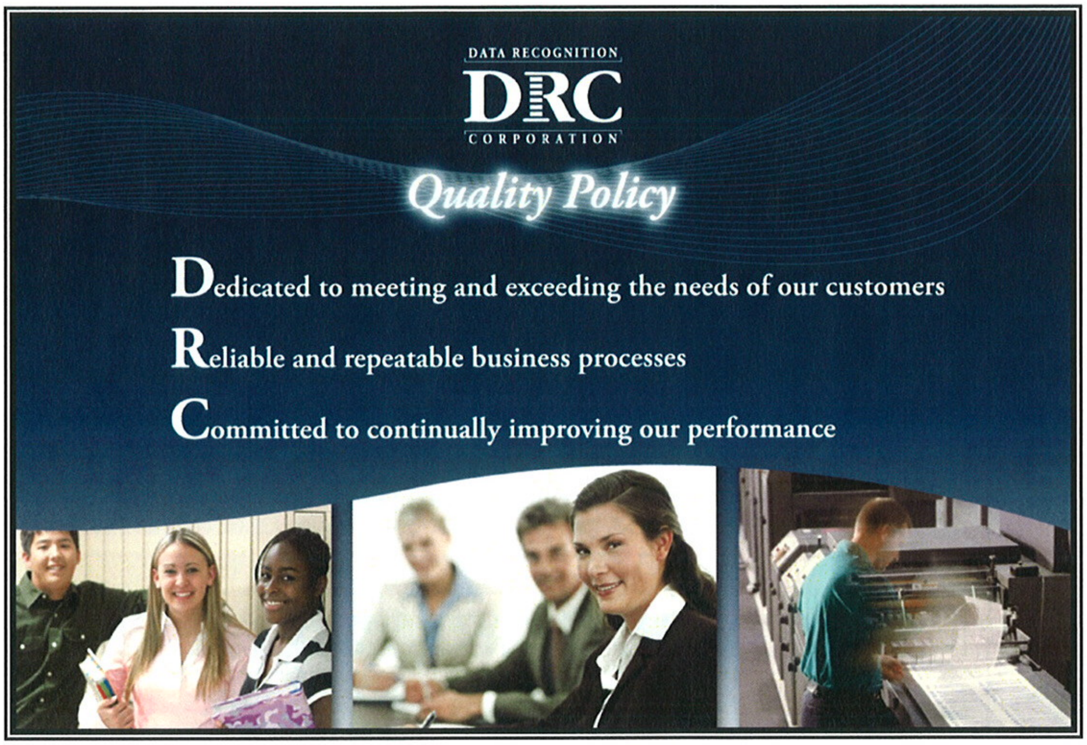
Please see *Subheading 6, Scoring,* for a more detailed discussion of DRC's scoring processes and associated quality management procedures and *Subheadings 8.a.v. and 8.a.vi.* for more information regarding our data file management procedures.

iv. Non-Scannable Materials Report

DRC will provide NDE with a report that will detail any damaged materials that could not be scanned. The report will list the number of materials that could not be scanned, as well as summarize any problems noted during materials return/check-in, processing, and/or scanning. Reports will be produced based on information from an error log maintained by Project Management. This report could be used to assist DRC and NDE in improving the instructions in the Test Administrator Manuals and Principal/Test Coordinator Manual, as well as information shared in the test administration training workshops.

DRC's Quality Management System

DRC is passionate about providing quality products and services to our clients and recognizes that quality processes are critical elements of our business. Quality at DRC is being taken to world-class levels, providing us with yet another competitive advantage. Figure 4–54 displays our Quality Policy.



DATA RECOGNITION
DRC
CORPORATION

Quality Policy

Dedicated to meeting and exceeding the needs of our customers

Reliable and repeatable business processes

Committed to continually improving our performance

Figure 4–54. DRC's Quality Policy

With 30 years of successful student achievement testing, we have developed and refined our quality system to ensure the highest levels of customer satisfaction and quality. **At DRC, quality is both a program and an overall approach to business.** Our Quality Management System is focused on defining and implementing critical quality control processes to ensure products and services delivered to our clients meet and exceed their requirements. This extends to our relationships with other vendors and partners.

At DRC, quality is a commitment to excellence and is achieved by teamwork and the process of continuous improvement. Quality principles are infused into everyone's roles within our organization. We are dedicated to being the quality leader in the industry and are confident our products and services will exceed NDE's expectations. The focus of our Quality Management System is to define and implement quality control processes and embed them throughout all aspects of our projects. DRC has developed our quality approach using the guidelines listed in the *SCASS/TILSA Quality Control Checklist for Processing, Scoring, and Reporting*. Our Quality Management System is illustrated in Figure 4–55.

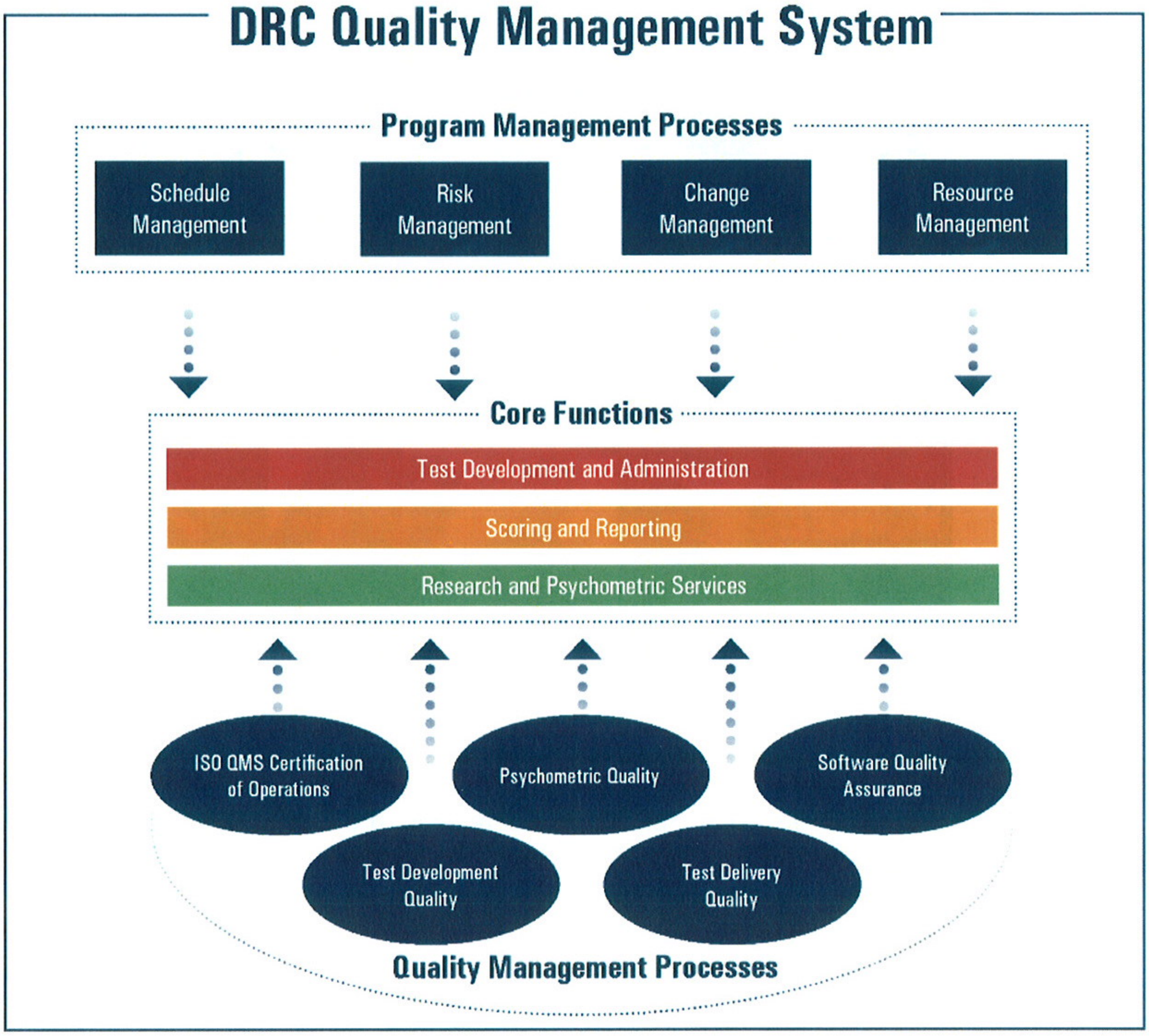

DRC Quality Management System
Program Management Processes
Schedule Management
Risk Management
Change Management
Resource Management
Core Functions
Test Development and Administration
Scoring and Reporting
Research and Psychometric Services
ISO QMS Certification of Operations
Psychometric Quality
Software Quality Assurance
Test Development Quality
Test Delivery Quality
Quality Management Processes

Figure 4–55. DRC's Quality Management System

Quality begins with the attitude that a task must be done right the first time. DRC staff members take great pride in their work, and their products reflect that pride. As demonstrated on all current DRC contracts, we understand the tasks that are necessary for successful assessment programs. DRC believes in reasonable and sensible approaches to problem solving. We pride ourselves on our creativity and our ability to anticipate problems, as well as our genuine affinity for discovering multiple solutions to difficult issues. DRC is eager to develop and maintain a mutually beneficial relationship with NDE.

ISO 9001 Certification

In today's world, customers continue to evolve their wants and needs. They are asking us to be more dynamic, flexible and cost efficient in meeting their requirements than ever before. This places a tremendous amount of importance on our processes to meet these needs in a reliable, repeatable fashion.

That is why DRC made the decision to attain ISO 9001 certification in 2007. ISO 9001:2000 is an internationally recognized quality management standard that defines a set of core quality requirements an organization must comply with. Some of the requirements in the ISO 9001:2000 standard include:

- A set of procedures that cover key processes within a business.
- Monitoring manufacturing and business processes to ensure they are producing quality products and services.
- Keeping proper records.
- Checking outgoing product for defects, with appropriate corrective action where necessary.
- Regularly reviewing individual processes and the quality system itself for effectiveness.
- Facilitating continual improvement customers expect.

DRC is currently ISO 9001:2000 certified in three major areas of the company:

- Document Services (Project Management, Publications, Pre-Press, Printing, Bindery, Inserting, and Purchasing).
- Educational Operations (Distribution, Logistics, Materials Processing, Warehousing and Document Scanning).
- Woodbury and Minnetonka, Minnesota and Cincinnati and Columbus, Ohio Scoring Centers.

External validation from a third party is required for a company to become ISO 9001 certified. An organization known as a "registrar" evaluates whether DRC is meeting the criteria of the ISO 9001:2000 standard within our quality management system. These "audits" are conducted twice annually.

The scope of our ISO 9001:2000 registration is based on a "business process", rather than a "functional" approach that many companies apply. Embedding the ISO 9001 standard has enhanced an already strong foundation of business process controls that has been DRC's hallmark for many years.

Our ISO 9001 certification process is led by **Mr. Niall Finn, Director of Quality for DRC's Operations.** Mr. Finn has extensive hands-on quality management experience in various manufacturing environments. As the senior quality leader responsible for leading the implementation of ISO 9001 Quality Management

System certification across all DRC operational areas, he will continue to oversee the plan to expand the scope of our certification to other areas of the company, while contributing his expertise to our quality standards and systems already in place.

Quality Project Management and Planning

For the success of the program, NDE's requirements, goals, and constraints must be thoroughly understood, documented, and communicated. These critical activities are the foundation of DRC project management activities. **Ms. Patricia Johnson, NeSA Project Director**, will be responsible for the administration of the overall quality process. Problem-reporting procedures will be strictly followed to ensure immediate action is taken to resolve any issues.

A primary factor in DRC's continued success in providing error-free services to clients is our company-wide dedication to quality.

DRC's **Vice President of Quality, Ms. Lisa Peterson-Nelson**, will also carefully audit the project delivery process for the NeSA. She is currently directing the enhancement of DRC's key work processes for delivery of products and services to clients. Ms. Peterson-Nelson has over 19 years of experience in quality process improvement. She worked for more than a decade in senior positions in quality process management for two different Fortune 500 companies. She has been with DRC since 2001.

We will provide NDE with the required evidence that our quality inspections, processes, system tests, and policies are followed. In addition, DRC will also provide NDE with a Quality Control Manual at the end of each contract year detailing the quality procedures used throughout all phases of the project. The manual will be updated yearly and will include any changes in processes or procedures.

To ensure the success of the NeSA, we will proactively manage risks, such as programmatic, technical, cost, and schedule risks. Ms. Johnson will function as the risk manager by working with other members of the project team to provide the work breakdown for the project, develop detailed scope of work agreements, and design of a formal risk management matrix specific to the project. Ms. Johnson will schedule and oversee risk reviews, in conjunction with the project team and NDE. Ms. Lisa Peterson-Nelson, DRC's Vice President of Quality, will also provide support to the risk management process, providing an additional level of program security.

NDE necessitates a partner that is flexible, innovative, and prepared to manage change. At DRC, change management is a critical management discipline. Our change management process is used to control and manage size, effort, cost, and schedules. Because change can occur at any time, we have implemented activities in our process to identify change, control change, and ensure change is properly implemented and reported to groups who are affected. NDE can be assured that DRC will thoroughly evaluate each requested change and perform a detailed impact and risk analysis. We will provide NDE with our recommended implementation plan and clearly outline any schedule or cost impacts.

Quality Control Process Overview

Our Project Delivery Quality Control process begins with the contract award and ends with the distribution of all required deliverables. Quality control checkpoints are in place at all stages of processing. Our proven quality framework is an integral part of ensuring accurate and timely delivery for our clients. We will provide NDE with the required evidence that our quality inspections, processes, system tests, and policies are followed.

Our Project Delivery Quality Control process is illustrated below in Figure 4–56.

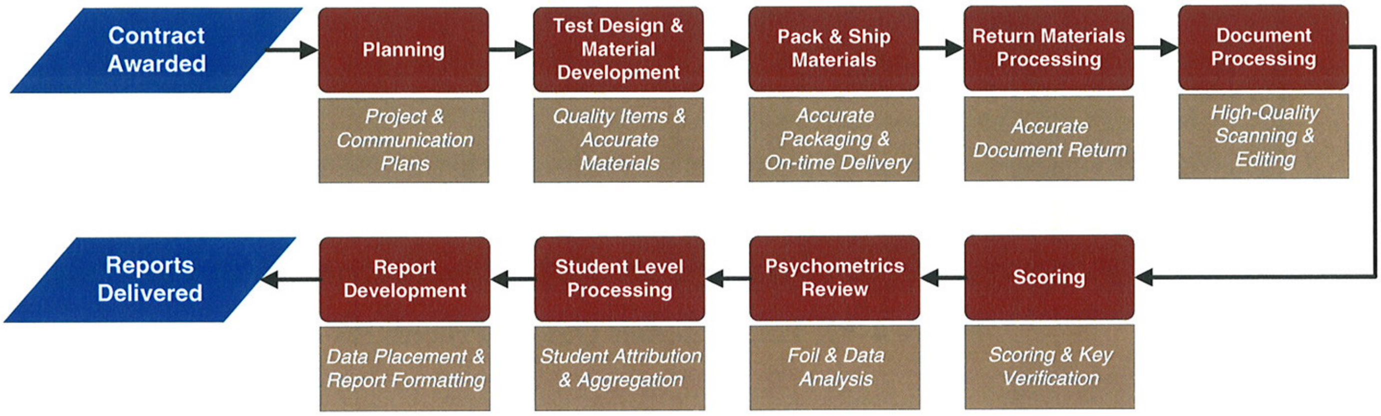

Contract Awarded
Planning
Project & Communication Plans
Test Design & Material Development
Quality Items & Accurate Materials
Pack & Ship Materials
Accurate Packaging & On-time Delivery
Return Materials Processing
Accurate Document Return
Document Processing
High-Quality Scanning & Editing
Scoring
Scoring & Key Verification
Psychometrics Review
Foil & Data Analysis
Student Level Processing
Student Attribution & Aggregation
Report Development
Data Placement & Report Formatting
Reports Delivered

Figure 4–56. Project Delivery Quality Control Process

Test Materials Quality Procedures

DRC understands that state departments of education require error-free materials; we take great pride in the excellence of the state testing publications we produce on behalf of our department of education clients. NeSA testing-related materials will be **produced through DRC's ISO 9001-certified Document Services Division** that incorporates our Document/Graphics Design Group and our complete in-house Printing Department. This incorporation of resources gives DRC a unique capability to customize our processes to address the requirements of each of our clients within restricted parameters and rigorous timeframes.

In addition to ISO 9001 certification, DRC has earned **Quality Level II status from the Government Printing Office (GPO),** which is the second highest status that can be awarded (Quality Level I status is reserved for printers who produce bound books, four-color varnished promotional pieces, or other similar materials).

DRC collaborates with each of our clients to maintain a program style guide to ensure consistent application of preferences and expectations across all program materials. Documents printed at DRC are printed to exacting specifications to guarantee the highest possible data integrity for OMR, OCR, and Imaging machines.

DRC follows a meticulous set of internal quality standards to ensure high-quality printed products for its clients. DRC assures NDE of our commitment to produce accurate test materials. Our quality procedures for production of test materials are highlighted in Figure 4–57.

- **Publishing and Editing Review—**DRC staff will perform a three-way review of all project materials. This process includes multiple group checks of answer keys to verify accuracy. After this internal review, assessment materials will be forwarded to NDE for review and approval.
- **Taking the test—**Staff will take the actual tests to ensure that all items and passages perform as planned.
- **In-house printing—**DRC's in-house printing department will print scannable materials based on predetermined specifications for quality and accuracy. External printing companies hired to print nonscannable forms will need to guarantee DRC the highest level of quality. DRC Project Management will review the vendor's quality plan.
- **Multiple checks—**DRC Project Managers and Print Procurement Specialists will routinely conduct meticulous quality checks during the printing process to see that all requirements for printed materials are met.

Figure 4–57. Test Materials Quality Procedures

Packaging, Shipping, and Materials Return Quality Procedures

Accurate packing, shipping, and collection of test materials are critical for districts and schools to successfully administer the NeSA. DRC is proud of our quality excellence in this area and are committed to upholding that level of excellence for NDE. Our quality procedures for packaging and shipping and materials return are summarized in are summarized below in Figure 4–58 and Figure 4–59, respectively.

- **Detailed instructions—**Based on contract requirements and specifications, detailed Scope of Work Agreements (SOWAs) will be established by the DRC Project Managers working in conjunction with our Operations staff. The SOWAs will be available for NDE review at each step of the process.
- **Walkthroughs—**The Project Management team will conduct a walkthrough of the assembly process prior to each shipment to check that all procedures are precisely followed.
- **On-going monitoring—**The Director of Materials Operations and the Logistics Manager will monitor the materials assembly area and report any irregularities to Project Management.
- **Secondary checks—**Our Operations staff will perform secondary checks on all packing lists and boxes will be sealed for shipping.
- **Easy identification—**All district and school shipping labels will be quality checked to prevent materials going to the incorrect location. Site labels on each box will be compared to the shipping address label and matched for accuracy.
- **Traceability—**Shipping carriers used have online, traceable distribution systems to track all materials.

Figure 4–58. Packaging and Shipping Quality Procedures

- **Tracking of boxes—**Upon receipt of materials at DRC, all returned boxes will be scanned in through our automated Box Receipt System. Quality control reports are generated to compare materials received against the shipper's manifest and the district counts. Materials return information will be reported to NDE on a daily basis.
- **Tracking of test materials—**After box receipt, test materials will be separated for processing using DRC's Operations Materials Management System (Ops MMS). Any discrepancies in expected counts of materials based on original packing will be reported to Project Management for resolution.
- **Missing materials reports—**DRC will generate missing materials reports, which will be available for NDE to review. After all materials have been checked in and discrepancies have routed for resolution, a final report will be generated for NDE.
- **Communication—**DRC's Project Management staff will communicate with NDE regularly during the entire materials receipt process to discuss any concerns or issues.

Figure 4–59. Materials Return Quality Procedures

Scanning Quality Procedures

DRC's image scanning and handscoring system was designed and built to work for all DRC imaging projects. Having a common scanning and handscoring system and platform eliminates the need for significant software development efforts to scan and score new projects. If enhancements are required for a project, the Imaging Information Systems department follows the proprietary DRC software development methodology to complete development. This methodology outlines the standard deliverables for each phase of the development lifecycle (analyze, build, test, implement). Prior to implementation, all enhancements are reviewed and verified by the Software Quality Assurance department (SQA).

DRC is committed to embedding quality throughout every aspect of our software design, development, and quality assurance processes, ensuring 100% accuracy in our scoring and reporting systems. **Ms. Karen Olsen, Senior Director of Information Systems Software Quality Assurance**, will oversee all software quality assurance activities for the NeSA. She has led the software quality assurance initiatives for DRC's Corporate Information Systems departments for over six years. With more than 12 years of experience in the software quality assurance field, Ms. Olsen has expertise in the implementation of exacting software quality assurance procedures throughout all phases of a project.

The Software Quality Assurance staff will apply industry-standard software quality assurance methodologies throughout the program. DRC quality plans will be developed and will be available for NDE's review, if desired. Software Quality Assurance staff follow our project delivery quality control process and adhere to the quality control checkpoints for processing, scanning, and editing, described by the State Collaborative on Assessment (SCASS) on Technical Issues in Large-Scale Assessments (TILSA). Our proven Software Quality Assurance standards and procedures are directly aligned with the Capability Maturity Model (CMM) from the Software Engineering Institute (SEI). DRC's scanning quality procedures are highlighted below in Figure 4–60.

- **Test Decks—**DRC will process test decks configured for the NeSA through the production systems.
- **Calibration—**Daily calibration and scanner cleaning processes will be conducted to ensure read level consistency.
- **Standard edit processes—**Every scannable document will be processed through edit programs to detect potential errors (double marks, smudge marks, omits, etc.)
- **Multiple reviews—**The Document Processing Supervisor will conduct a review of the entire first batch prior to full production to ensure error-free processing.
- **Quality control reports—**Daily quality control reports for each editor will be reviewed by the Document Processing Supervisor to monitor the accuracy of the online editing process.

Figure 4–60. Scanning Quality Procedures

Scoring Quality Procedures

DRC understands the activities and coordination required for data processing and scoring of the NeSA and has the proven experience and capabilities to score the tests accurately. DRC brings many years of valuable and accurate scoring experience spanning across programs such as Alabama, Alaska, Louisiana, Pennsylvania, and South Carolina.

We will prepare and refine the requirement documents for the scoring of answer sheets well in advance of the receipt of test materials. These specifications will contain detailed scoring procedures, along with the procedures for determining whether a student has attempted a test and whether they should be included in statistics and calculations for computing summary data.

The requirement documents will be completed and reviewed with NDE. After all changes and edits have been made, the final requirement documents will be sent to NDE for final approval.

DRC's strict quality procedures can assure NDE accurate scoring. **We are prepared and accustomed to handling programs with multiple forms at various grade levels and/or content areas** and have built-in solid checkpoints and reviews throughout the entire scoring process. Standard quality inspections will be performed on all data files, including the evaluation of each student data record for correctness and completeness prior to report generation. Student results are kept confidential and secure at all times.

Score Key Quality

The integrity of item, form data, and score keys will be evaluated in several ways. Similar to our score key validation procedures used on other assessment programs, we will leverage our established, documented process to ensure all score keys are accurate. Test development specialists, psychometric staff, and software quality assurance analysts will check the score keys through a series of validation procedures at varying junctures. Our score key quality procedures are summarized below in Figure 4–61.

- **Verify for accuracy**—Score keys will be verified for accuracy based on multiple reviews by test development specialists, psychometric staff, and software quality assurance analysts. All item data and score keys will be reviewed and approved by each group prior to scoring NeSA tests.
- **Take the test**—Multiple staff with specific content knowledge will take each form of the test and compare their results against the score keys on the test maps. The score keys and strand information will again be verified during this step.
- **Score key file import**—DRC will import the approved keys received into our scoring system. Once the keys are successfully imported, software quality assurance staff will re-verify the keys used by the scoring engine.
- **Database accuracy**—All items will be scored in the system using the correct and incorrect item distractors. The database will be validated to make certain the distractor captured in scanning was saved correctly and that the item was given a correct or incorrect answer.
- **Automated system checks**—The scoring engine has automated system checks built-in to validate score keys and proper merging of multiple-choice and constructed-response items. Additionally, the software quality assurance team performs independent checks on this data.

Figure 4–61. Score Key Quality Procedures

Data File Quality Control

DRC understands the critical nature of scoring large-scale assessments. Our systematic approach will ensure successful scoring and 100% accuracy. DRC has the thorough understanding of the requirements needed to monitor, score, and effectively analyze the data for the NeSA.

All **data file development for the NeSA will be done in close association with NDE** to ensure requirements are met. Each data file produced will be **quality checked** for accuracy and completeness a **minimum of three times** by DRC's Software Quality Assurance Analysts and Project Management staff against NDE-approved layouts, specifications, and processing rules.

Online Systems Quality Control

DRC is proud of the web-based systems that we have created in conjunction with many state departments of education over the years. Our commitment is to deliver high-quality content and error-free, reliable web-based systems to NDE and Nebraska educators and students. Recognizing that quality is the most critical element of our business, we have developed and refined our quality system to ensure the highest level of quality and customer satisfaction will be provided to our clients.

DRC's **quality assurance staff will monitor development of all web-based systems.** Figure 4–62 presents our quality criteria for web-based systems.

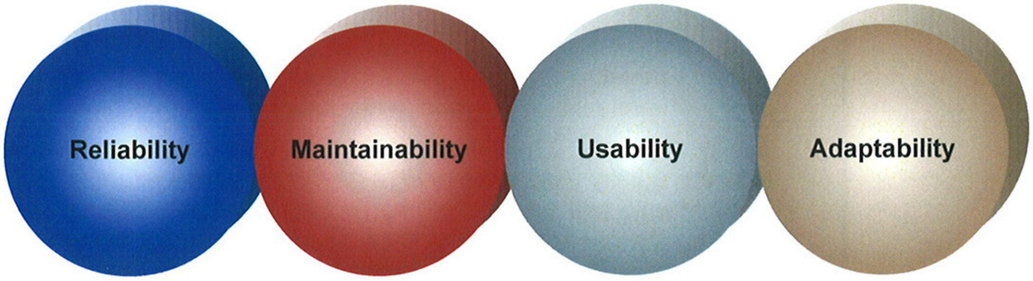


Figure 4–62. Web-Based System Quality Criteria

DRC staff will follow our standard project delivery quality control process and adhere to quality control checkpoints described by the State Collaborative on Assessment (SCASS) on Technical Issues in Large-Scale Assessments (TILSA). Our proven quality standards and procedures are directly aligned with the Capability Maturity Model (CMM) from the Software Engineering Institute (SEI).

Online System Functionality, Security, and Performance Quality

The functionality, security, and performance requirements of the test delivery system will be clearly documented. **All system and processing requirements will be documented based on NDE's decisions**. These documents will serve as the system's scope and will be used to validate overall functionality.

Prior to the release of the test delivery system, the quality assurance staff will perform full system-level tests in an independent test environment that simulates the production configuration. The system will be tested on all supported computer platforms and browsers. The system-level tests will include comprehensive assessments on functionality, usability, reliability, security, and overall performance. Each web page, link, item, and image will be verified to ensure it is displaying properly, following Graphical User Interface (GUI) standards, and functioning as designed. System content will also be validated for accuracy during this process. Our online system functionality, security, and performance quality procedures are presented in Figure 4–63.

Once approved, the system will be released to NDE, if desired, for final verification. Upon final approval, the system will be moved to the production environment where it will again be verified to assure it is ready for use by schools and students.

- **Unit testing**—System features will be subjected to functional testing by the software development staff. At this stage, issues can be detected and corrected prior to the release to the quality assurance staff.
- **System testing**—The system will be subjected to system-level testing by software quality assurance staff. At this stage, the system will be validated against requirements and subjected to full functional testing. This process includes verifying system accessibility, links, scoring, reporting, security, and performance. Issues can be detected and corrected prior to the final release.
- **Editorial review**—A multi-step editorial review of all item computerized displays, including graphs, charts, illustrations, and tables will be performed.
- **Install/uninstall testing**—Installation procedures, updates, and patches will be fully tested prior to releases.
- **Load testing**—Simulation of heavy loads on the system will be performed to confirm that the solution will meet performance expectations.
- **Security testing**—Extensive tests will be performed to ensure security requirements are being met on the system and user access is limited to the appropriate security level.
- **Database accuracy**—Quality assurance staff will perform extensive tests to ensure all data captured in the test delivery system is stored in a secure database environment.
- **Independent Department review**—The system can be provided to NDE for validation prior to the release to schools and students.

Figure 4–63. Online System Functionality, Security, and Performance Quality Procedures

Online System Data Validation

In addition to score key validation, experts will conduct related quality checks on the system data. Quality control checks (see Figure 4–64) will be performed throughout the system-level testing, including checks of imported and reported data results, to ensure the integrity of the data.

- **Duplicates—**The system will be checked for duplicate records and items.
- **Scored data—**Quality checks will be performed on the data to ensure that test scores have been computed correctly against the score keys and scoring requirements.
- **Data standards—**Standard database and data naming conventions will be established and used.
- **Database accuracy—**Quality assurance staff will perform extensive tests to ensure all data captured in the test delivery system is stored in a secure database environment.

Figure 4–64. Online System Data Quality Procedures

Quality: A Corporate-Wide Value at DRC

As described above, DRC has in place the necessary quality control processes—from initial project planning through the delivery of final reports—to successfully develop and administer large-scale assessment programs. Through our Quality Management System, DRC feels confident in guaranteeing the accuracy and on-time delivery of our large-scale assessment projects. We look forward to providing these high-quality services to NDE and the State of Nebraska.

6. Scoring

a. Key Verification

DRC understands the activities and coordination required for data processing and scoring of the NeSA program. **Delivering assessment results on time and without error is critical.** DRC scores over 5 million student answer documents on an annual basis and has successfully matched multiple-choice, constructed-response, and online-response data without any reported errors. Our scoring and reporting systems have quality procedures integrated throughout, including both automated and manual inspections, to ensure data accuracy. DRC's experience and expertise will directly contribute to the successful processing and reporting within the prescribed time limits. **Ms. Karen Olsen, DRC's Senior Director of Software Quality Assurance**, will oversee all aspects of the process for merging scores and will ensure timely delivery of NeSA results.

All student answer sheets returned to DRC will be scored. Multiple-choice items and demographic information will be image scanned and the original scanned data will be converted into a master student file. Record counts will be verified against the counts from the Document Processing staff to ensure all students are accounted for in the file. Additionally, a detailed review of the error-tracking log will be performed to ensure any discrepancies are resolved before proceeding with the scoring routines.

The multiple-choice items will be scored against the appropriate answer key, indicating correct and incorrect responses. In addition, the student's original response string is stored for data verification and auditing purposes. We will prepare and refine the requirements documents for the scoring of answer sheets well in advance of the receipt of test materials. These specifications will contain detailed scoring procedures, along with the procedures for determining whether a student has attempted a test and whether they should be included in statistics and calculations for computing summary data. DRC will ensure that all answer keys have been approved by NDE and verified for accuracy prior to the scoring of any student responses. Student scale scores and achievement levels will be determined prior to the production of final data files and reports.

DRC's strict quality procedures will result in accurate scoring. We are prepared and **accustomed to handling programs with multiple forms, assessments, and testing modes** at each grade level and have built-in solid check-points and reviews throughout the entire scoring process. **We have not encountered any situations where student scores have been matched incorrectly using our process and established quality control procedures.**

Once the scored master student file is deemed 100% accurate, DRC's Psychometrics staff will perform additional detailed analysis on the data files prior to NDE's review and approval process. Standard quality inspections will be performed on all data files, including the evaluation of each student data record for correctness and completeness. Student results are kept confidential and secure at all times. Figure 4–65 illustrates DRC's scoring quality process.

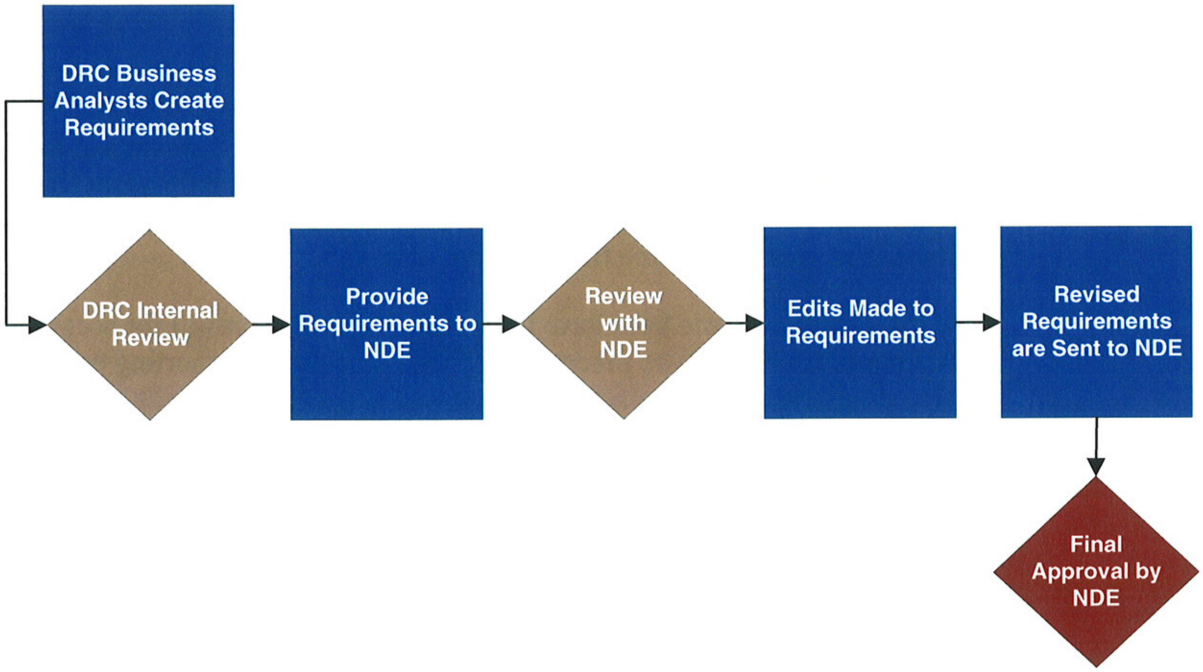

DRC Business Analysts Create Requirements
DRC Internal Review
Provide Requirements to NDE
Review with NDE
Edits Made to Requirements
Revised Requirements are Sent to NDE
Final Approval by NDE

Figure 4–65. DRC Scoring Quality Processes

Software Quality Assurance Testing for Data Analysis and Verification

To provide NDE with the highest level of accurate test results, **DRC will conduct a thorough evaluation of all scored data.** File formats and data elements will be validated against client-approved layouts, specifications and processing requirements. Detailed test scripts will be executed to confirm accuracy. Some of the quality verification steps include:

- Verification of answer keys/test maps
- Raw scores
- Raw-to-scale score conversions
- Scale-score comparisons to performance achievement levels
- Disaggregated data
- Processing rules for individual student and summary level data

The quality assurance steps involve processing sample student records through the data processing and scoring system. Each student's data record will be carefully reviewed and evaluated to ensure it was **scored with 100% accuracy.**

To reduce the risk of human error, our Software Quality Assurance programmatic test routines will be used to thoroughly evaluate each student's data record that will be produced for use in final data files and reports.

Key Verification

DRC will ensure the accuracy of the answer keys. At least two DRC staff with content-specific expertise will take the test and compare their answers to the answer keys on file for each test item. Test-takers will also verify academic standard information. Quality Assurance staff will compare the processing file against the answer key source file to ensure accuracy.

Once the multiple-choice keys have been analyzed and approved for accuracy, clean, edited batches will be processed through scoring and reporting programs. Scoring programs will contain answer keys and academic standards categorizations for each item. Items will be scored as right, wrong, omitted, or double-gridded. After scanning and scoring, DRC Quality Assurance staff will verify values for multiple-choice items.

Online Test Key Verification

To ensure scoring key accuracy of the online NeSA assessments, CAL staff members will take each NeSA grade level and content area test and compare their answers to the answer keys in our NeSA test directory that have been verified and approved by NDE.

Additionally, CAL will perform a statistical validation of the answer keys as 2000 student responses per subject or domain become available. All suspect items are immediately referred to content specialists for further review and verification. Any questionable item keys will be sent to NDE for review.

b. Merging Online and Paper/Pencil Results

Students' paper/pencil multiple-choice responses will be scored using DRC's proven scanning and scoring procedures described above. Scores for a student's online responses will be securely transferred from the CAL system to DRC for student level processing and reporting. Both paper/pencil and online scores will be incorporated into the master scoring database for each administration prior to analysis being performed.

Scores for a student's multiple-choice responses for paper/pencil tests will be systematically matched to online multiple-choice scores (as appropriate) by a unique document ID (lithocode) and/or a series of criteria (e.g., NDE Student ID number, first/last name, district/school, birthdate) determined in collaboration with NDE during the requirements gathering process for scoring and reporting. This process allows DRC to create a **single, accurate, reliable data record for each student assessed** by linking all score and demographic data for a specific student, including data and scores collected during scoring of paper/pencil and online multiple-choice response items.

DRC's strict quality procedures will result in accurate scoring. We are prepared and **accustomed to handling programs with multiple forms, assessments, and testing modes** at each grade level and have built-in solid check-points and reviews throughout the entire scoring process. **We have not encountered any situations where student scores have been matched incorrectly using our process and established quality control procedures.**

Once the scored master student file is deemed 100% accurate, DRC's Psychometrics staff will perform additional detailed analysis on the data files prior to NDE's review and approval process. Standard quality inspections will be performed on all data files, including the evaluation of each student data record for correctness and completeness. Student results are kept confidential and secure at all times.

Figure 4–66 outlines DRC's scoring process, including merging student data from multiple scoring sources.

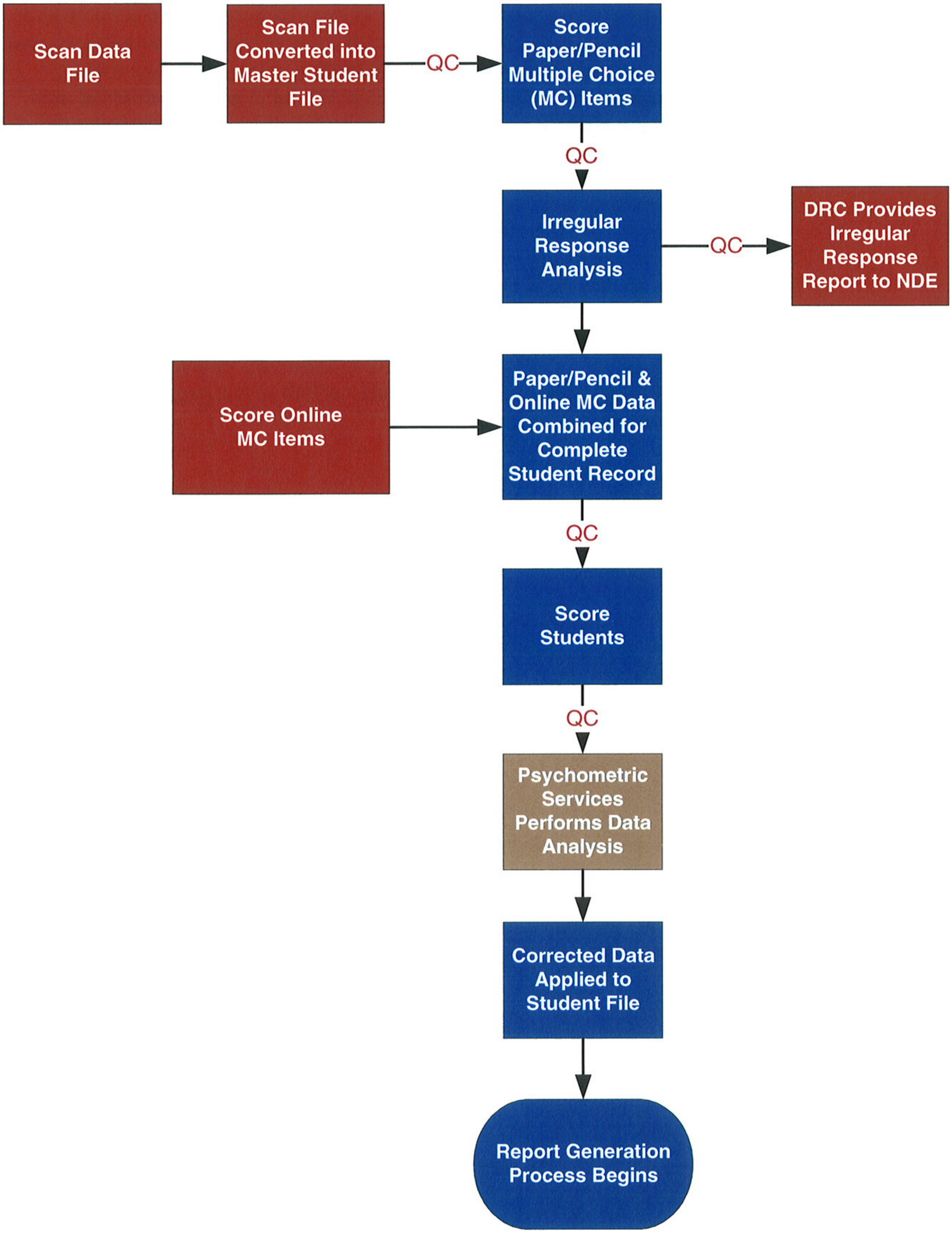

Scan Data File
Scan File Converted into Master Student File
QC
Score Paper/Pencil Multiple Choice (MC) Items
QC
Irregular Response Analysis
QC
DRC Provides Irregular Response Report to NDE
Score Online MC Items
Paper/Pencil & Online MC Data Combined for Complete Student Record
QC
Score Students
QC
Psychometric Services Performs Data Analysis
Corrected Data Applied to Student File
Report Generation Process Begins

Figure 4–66. Process for Merging Student Data and Scores

c. Irregular Response Report

Following each test administration, DRC will provide NDE with a report documenting irregular responses, including blank answer documents, excessive item non-response, excessive multiple marks, and erasure data (wrong-to-right). We will provide this report broken down by district and school. DRC and NDE will collaborate to determine the indicators of irregular response.

DRC staff will support NDE as it addresses any resulting issues, including potential test security breaches. Our support could include documentation and/or additional data analysis. Costs for additional analysis would be provided upon request.

7. Analysis

DRC will conduct all analyses necessary to ensure that tests meet standards of technical quality and report meaningful results for the student, school, district, and state for the Nebraska assessment. DRC achieves psychometric excellence by assuring that all practices and procedures meet professional measurement standards as outlined in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

During each year of the contract, DRC will conduct analyses necessary to support:

- Test development for test items developed by NDE.
- Test construction.
- Scoring.
- Standard setting and validation activities.

Upon award, DRC's research staff members look forward to working with NDE to define a scope of work to conduct secondary analyses related to security, data interpretation, policy formation and administrative planning. DRC has several options available for large-scale assessment clients, including erasure analysis, Rasch residual analysis, as well as AYP data analysis. *Appendix F* contains a more detailed discussion of a possible approach to one aspect of secondary analysis (Data Forensics). DRC will work with NDE to implement this or other secondary analyses that might prove both useful and influential in regard to policy formation and administrative planning.

a. Calibration and Scaling

i. Calibration of Test Items

DRC proposes the Rasch Measurement Model to manage the testing and assessment process. There are several benefits to using the Rasch Measurement Model in large-scale assessment (Smith & Smith, 2004; Mead, 2008):

- It is relatively simple to apply, which aids communication and allows tight reporting schedules.
- It provides an interval scale of measurement, which permits direct comparison of students and items.
- It separates the information relevant to the measurement from the error terms, which facilitates detection and diagnosis of irregularities.

In order to derive data that he considered worthy of the name *measurement*, Georg Rasch (1960) reasoned that only one person parameter (*ability*) and one item parameter (*difficulty*) can govern the interaction between the person and the item. If the person has more ability than the item has difficulty, the person is expected to answer correctly. If the person has less ability than the item has difficulty, the person is expected to answer incorrectly, regardless of any other characteristics of the person or item.

This line of reasoning led to the *simple logistic model*, which has several closely related and very useful properties:

- *Separability* of the model parameters (Rasch, 1960).
- *Sufficient statistics* that do not involve the parameters (Andersen, 1977).
- *Specific objectivity*, sometimes called person-free calibration and item-free measurement (Wright, 1968).
- *Simplicity*, which allows ready explanation and understanding of the measures (Wright & Stone, 1979).

Specific objectivity means that the estimation equations for ability do not involve the difficulty parameters, and the equations for difficulty do not involve the ability parameters. In practical terms, this means that students can be ordered along the measurement continuum by their number correct scores and that items can be ordered along the same continuum by the number of correct responses to the item. No other information is necessary and anything remaining in the data can be used for control of the model. *Specific objectivity* is the cornerstone of the Rasch family of measurement models (Wright, 1980).

DRC is confident that the processes it employs in those phases will adequately satisfy the demands of the model and permit NDE to enjoy the advantages of this model for effective reports, quality control, and timely turnaround of results.

Item Calibration

The multiple-choice items (MC) will be calibrated using the familiar form of the Rasch model (Rasch, 1960; Wright & Stone 1979; Andrich, 1988; Fischer & Molenaar, 1995; Smith & Smith, 2004). The Rasch model applicable to dichotomously scored items, scored right or wrong conditional on the ability and difficulty, can be expressed in the most familiar form of the model:

$$\Pr(right \mid \beta_n, \delta_i) = \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}}.$$

The probability of success for a person with ability β_n on an item with difficulty δ_i is determined by the difference between the ability of the student and the difficulty of the item.

Software

Joint-maximum-likelihood estimation of items will be accomplished using WINSTEPS (2008). This calibration software is commercially available and widely used in the testing industry. The capabilities of the WINSTEPS program will be utilized to assess unidimensionality, item interdependence, and other deviations from the model, as well as item calibration and ability estimation. The program has several options for the exploration of the person-item residual matrix (Mead, 1976, 2008; Smith, 2000; Ludlow 1986).

An important consequence of these models is that the number of correct responses to a given set of items is a sufficient statistic for estimating person ability. As a result, each person with the same raw score will be assigned the same estimated ability.

The DRC Psychometric staff is experienced with this software for Rasch analysis. Members of the staff have been instrumental in the development of the model and its application for several decades.

ii. Translating Student Composite Scores to Reporting Scales

DRC is proposing the Rasch measurement model to estimate the student proficiencies and to control the assessment process. The model provides straightforward algorithms to compute *ability* estimates on a unidimensional, equal-interval scale of measurement from the number correct scores. These algorithms are implemented in the WINSTEPS program and other readily available software.

The native Rasch scale of measurement, often referred to as *logits*, has many attractive measurement properties, but communication with the public is not one of them. Logits involve positive and negative values and two, three, or more decimal places. The scale score metric used for reporting is, in almost all cases, a simple linear transformation of the logits; *SS = a+b (logit)*, where b is large enough to avoid the need for decimals and a is large enough to avoid negative values.

Beyond the considerations just mentioned, the choice *a* and *b* are completely at the discretion of NDE. Any linear transformation will retain the measurement properties of the logit metric. The choice should attempt to create the most effective communication and interpretation of the results. Because there are exactly two parameters used in the logit to scale score conversion, exactly two aspects of the final scale can be fixed.

In many cases, the scale scores are fixed for important points (e.g., *Basic* and *Proficient* performance levels) at easily remembered values (perhaps, 250 and 500 or 100 and 200). It is then very convenient to compare any scale score to these benchmarks. In other cases, the state mean and standard deviation for the base year are set to round numbers, say 500 and 100 or 100 and 10. This facilitates a more normative interpretation[1]. An alternative is to set the mean for each grade in the base year to the grade times 100 plus 50 (e.g., the grade 3 mean in the base year would be 350.) This suggests (but does not actually create) a vertical scale.

DRC will work with NDE to establish the scale that best meets the needs of educators in Nebraska.

iii. Developing Scales to Report Subscore Results

State-wide, standards-based assessments are typically asked to perform two very different functions:

- System-level accountability, and
- Student-level diagnostics.

If it is appropriate to combine results from all items and subscales into a single score to determine performance level status for accountability purposes, there should be little or no diagnostic information in the subscores. If, on the other hand, students differ dramatically in the pattern of their subscale profiles, it is probably not appropriate to attempt to summarize them with a single score for accountability. In practice, the situation is rarely this dire; most students perform in a manner sufficiently consistent with a single dimension to justify the unidimensional assumptions while enough students deviate from this pattern to warrant reporting and analyzing subscale results.

There are three possible approaches to reporting subscale results: raw scores, scale scores, and pseudo-Z scores. DRC does not recommend using raw scores for this purpose. The only arguments in their favor are that users are familiar with them and believe, often mistakenly, that they understand how to interpret them. Differences in the number of items and in the difficulty of the items often make the raw score a misleading indicator of status. While reporting percent correct rather than number correct and including the state average or similar data as a frame of reference are attempts to address these deficiencies, they are not as effective as the application of a strong measurement model.

[1] Unfortunately, the (500, 100) scale would resemble SAT scores; the (100, 10) scale would resemble IQ scores. It is generally recommended that the scale selected not resemble an existing scale, although this is becoming more difficult.

The *specific objectivity* property of the Rasch model ensures that any selection of appropriate items will produce statistically equivalent estimates of the person's ability. Consequently, ability estimates can be made from any subscale and reported in the same metric as the total score estimate. Because they are in the same metric, the subscale estimates in the scale score metric can be compared directly to each other, to the total score estimate, to the item locations, to normative information, and to any other benchmarks that are helpful in the interpreting the results. Standard errors are also available to realistically access differences.

The subscale ability estimate for a raw score of *r* can be easily computed using the expression, usually iteratively,

1. $b_r^t = b_r^{t-1} + \dfrac{r - \sum_{i=1}^{L} p_{ri}^{t-1}}{\sum_{i=1}^{L} p_{ri}^{t-1}(1 - p_{ri}^{t-1})}$, where the summations are over the items included in the subscore and the initial estimate is either the total score estimate or ln {r/(L-r)}.

The asymptotic standard error for b_r is:

2. $se_r = \sqrt{\dfrac{1}{\sum_{i=1}^{L} p_{ri}(1 - p_{ri})}}$.

While from a psychometric perspective, reporting the total score and the subscore scale scores in the same interval scale metric is attractive, it is sometimes over-interpreted in the field. Even when standard errors are provided, there have been many instances where educators and policy makers have acted upon differences that could have arisen by chance. Confusion can also arise when the confidence intervals around subscores overlap the lines defining the performance levels.

Some of these problems can be mitigated if the subscore results are reported in a scale-free, standardized *pseudo-Z* metric. This removes the subscales from the scale score metric, but still permits, even facilitates, diagnosing strengths and weaknesses.

A simple calculation of pseudo-Z scores can replace the subscale ability estimates:

3. $Z_{vi} = \dfrac{r_v - \sum_{i=1}^{L} p_{ri}}{\sqrt{\sum_{i=1}^{L} p_{ri}(1 - p_{ri})}}$.

Expressions 1 and 3 have much in common: Z will tend to be large when the subscale ability differs dramatically from the overall estimate, with a positive value associated with an increase, and a negative value associated with a decrease. Compared to the scale score metric, the magnitudes of the Z's will be tempered so that less weight is attached to subscales and items that are far from the person's location. While it loses some of the attraction of the pure scale score metric, the Z-metric still allows identification of the student's strengths and weaknesses and is less susceptible to over-interpretation.

The strong measurement model also provides some protection from falsely reporting subscale differences that are simply due to random fluctuations of multiple pair-wise comparisons. When there are a number of subscales involved, there can be a large number of possible comparisons, which increases the risk of false positives. Before investigating possible subscale effects, there should be an overall assessment of the consistency of the entire set of effects for the student. Unless this assessment concludes the pattern is not random, it is not appropriate to attempt to interpret any subscale differences, although some individual differences could appear large.

There are two aspects to Rasch diagnostics: diagnosis *with the model* and diagnosis *from the model*. Either situation can be utilized to construct informative, individualized student and group reports that provide scaffolding to assist the user in interpreting the results.

Diagnosis *with the model* applies when all subscales give statistically equivalent results. Then the result can be interpreted by referring to the scale definition, which identifies the items and subscales that are relatively easy (i.e., below the student's level) and which are relatively difficult (i.e., above the student's level.) It is then appropriate to talk about which topics have been *mastered* and which are next to be tackled. This structure provides a clear understanding of how far the student has progressed and how far is left to go.

Diagnosis *from the model* occurs when the student performs inconsistently: surprisingly well on some subscales and surprisingly poorly on others. While it may not be clear what the student's overall status is, the diagnostic information from the subscale reports can help the frontline educators understand the student's needs, interests, and strengths.

Upon award, DRC staff will work with NDE to determine an approach to reporting subscore scale results that will provide the most effective communication and the most meaningful results for Nebraska educators.

b. Equating

i-ii. Equating Procedures

Rasch Equating

Angoff (1971) outlined three conditions that must be satisfied for equating to succeed:

- The test forms to be equated should measure the same ability (unidimensionality).
- The resulting raw score to scaled score conversion should be independent of the data used in deriving it and should be applicable in all similar situations.
- The equating should be symmetric, or the equivalent, regardless of which test form is designated as the base.

To succeed, in this sense, means that after equating, scores on the two test forms are interchangeable and a student's score may be compared to another's within or across years in an equitable and objective manner.

Test forms that conform to Rasch's principles are assured of satisfying Angoff's requirements. That being said, building forms to this standard is a challenging and on-going task. It requires the careful development of items to ensure the content of the items is consistent with the content standards, the curriculum, and the instruction. It also requires strong statistical controls to ensure that all items are equally valid and reliable instances of the underlying construct. Because of the strict requirements of the Rasch model, it is the ideal vehicle for providing these controls. Strict adherence to the model's requirements will be the guiding principle to develop sound measurement scales, to maintain consistent performance standards, and to facilitate comparable reporting across forms and across years.

The *specific objectivity* property of the model allows, once the sufficient statistics have been removed from the scores, the remaining data to be used for control and monitoring. The data should have no lingering influences dependent on the distribution of ability in the group who provide the calibrating sample, nor on which administration is considered. Any patterns related to years or groups will be dissected to determine how the differences arise. Like most forms of data analysis, the statistics will be used to call attention to problematic situations, but the substantive interpretation will require the collaborative efforts of DRC staff, educational specialists, NDE, and its technical advisory committees.

DRC is proposing a test design for operational assessments that will include:

- Core set of items,
- Linking (anchor) set of items, and

- Embedded field test items.

The exception will be in the first administration where no linking set is required.

In the future operational assessments, the linking set will not exceed 30 percent. DRC will choose the linking items from the previous year's assessment to place on the current year creating a year-to-year link. Considerations for the linking set of items include good content balance, breadth of difficulty, and good fit to the model.

As discussed in *Subheading 2.d., Content of Test Forms*, NDE is requesting one form be created annually using the previous years embedded field test items. There will also be one breach form constructed each year. The breach form is constructed in the same manner as the operational form. The breach form will be linked back to the previous administration. This link will comprise a different selection of linking items than the operational assessment, minimizing item exposure.

The standalone field test and the operational administration forms and the breach forms will then be parallel and spiraling will provide *randomly equivalent* samples that are adequate to place all items and forms on a common scale score metric. Information on this can be found in *Subheading 2.d.*

Pre-Equating

The test design DRC has chosen allows for both pre- and post-equating the assessments, in consultation with NDE. Pre-equating was chosen to achieve the rapid reporting that is desired by NDE and can accommodate all subjects once they have become operational and standards have been set. In the first two administrations, the field testing and first operational administration, which occur prior to setting the performance standards, the NeSA will be post-equated. No data are available prior to the field test, so post-equating is the only option.

The standalone field test data will be used to construct the first operational form. While forms should be interchangeable in content and difficulty (i.e., pre-equated), the final scale score metric will not have been established until after standard setting. Because the schedule required for standard setting will preclude the two-week reporting schedule, time will be available to conduct a post-equating analysis to further verify the field test results, calibrations, and equating.

This will provide a well defined and validated measurement scale. Beginning in the third year of implementation (i.e., the year after standard setting), each assessment can be pre-equated to facilitate the timely reporting of results.

It should be noted that the process of phasing in content areas at the rate of one per year has some implications for the reporting schedule. In year three, reading with pre-equating could be reported on the accelerated schedule. In the same year, math results need to wait for the standard setting. Upon award, DRC will discuss with NDE whether the two areas should be reported separately or on the math schedule.

Pre-equating forms places more emphasis on field test statistics and NDE can find more information on DRC's item evaluation analysis in *Subheading 7c*.

DRC's pre-equating procedure is based on the Rasch measurement model. Here, the one-to-one association between raw scores and ability estimates defines the raw to scale score conversions needed for scoring and reporting. For a given raw score (*r*) on a specific form of the test, ability is the value that makes the following equation true:

1. $$r = \sum E_{rj}$$

where E_{rj} represents an expected item score. For the multiple-choice items, each worth one point, this is the probability that a student will answer the item correctly, which is given by:

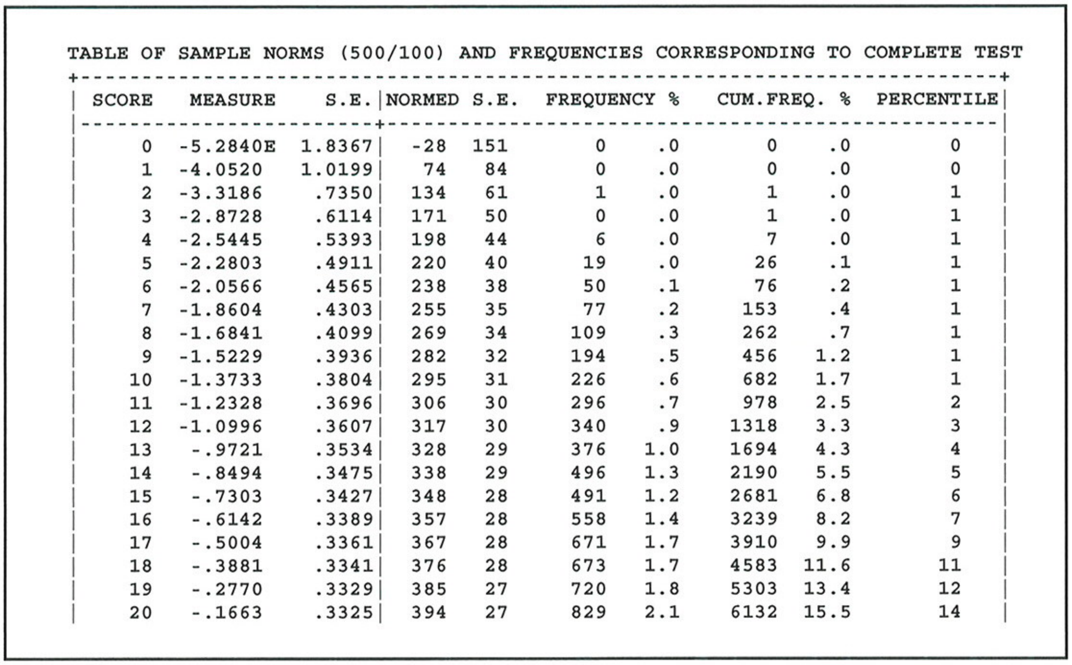
2. $$p(x_{vi} = 1 \mid \beta_v, \delta_i) = \frac{e^{\beta_v}}{e^{\delta_i} + e^{\beta_v}} = \frac{e^{\beta_v - \delta_i}}{1 + e^{\beta_v - \delta_i}}\ .$$

Forms of a test will be *equated* if all refer to the same measurement scale; two number correct scores will be *equated* if they refer to the same location on the measurement scale. Scale scores from equated forms can be analyzed, compared, and dissected without regard for which form or which administration generated the scores.

Pre-Equating Procedure

DRC will use WINSTEPS to generate the conversion tables. DRC psychometricians are very familiar with WINSTEPS output formats and have experience handling multiple output files programmatically. Psychometric staff will run WINSTEPS by anchoring all item difficulties and thresholds. WINSTEPS raw-to-measure conversions are completely model driven in such a 'fully anchored' run. Raw-to-measure tables will be obtained from the subsequent WINSTEPS output files. These files will then be checked using independent procedures by DRC's Psychometric Quality staff as discussed below.

A sample WINSTEPS scoring table is presented in Figure 4–67. The raw score to measure conversions are provided in the first two columns. The measure scores are linearly transformed to derive the NeSA scale score.



```
TABLE OF SAMPLE NORMS (500/100) AND FREQUENCIES CORRESPONDING TO COMPLETE TEST
+-------------------------------------------------------------------------+
| SCORE    MEASURE     S.E. |NORMED S.E.  FREQUENCY %    CUM.FREQ. %   PERCENTILE|
|---------------------------+---------------------------------------------|
|     0  -5.2840E  1.8367|  -28  151       0   .0        0   .0         0  |
|     1  -4.0520   1.0199|   74   84       0   .0        0   .0         0  |
|     2  -3.3186    .7350|  134   61       1   .0        1   .0         1  |
|     3  -2.8728    .6114|  171   50       0   .0        1   .0         1  |
|     4  -2.5445    .5393|  198   44       6   .0        7   .0         1  |
|     5  -2.2803    .4911|  220   40      19   .0       26   .1         1  |
|     6  -2.0566    .4565|  238   38      50   .1       76   .2         1  |
|     7  -1.8604    .4303|  255   35      77   .2      153   .4         1  |
|     8  -1.6841    .4099|  269   34     109   .3      262   .7         1  |
|     9  -1.5229    .3936|  282   32     194   .5      456  1.2         1  |
|    10  -1.3733    .3804|  295   31     226   .6      682  1.7         1  |
|    11  -1.2328    .3696|  306   30     296   .7      978  2.5         2  |
|    12  -1.0996    .3607|  317   30     340   .9     1318  3.3         3  |
|    13   -.9721    .3534|  328   29     376  1.0     1694  4.3         4  |
|    14   -.8494    .3475|  338   29     496  1.3     2190  5.5         5  |
|    15   -.7303    .3427|  348   28     491  1.2     2681  6.8         6  |
|    16   -.6142    .3389|  357   28     558  1.4     3239  8.2         7  |
|    17   -.5004    .3361|  367   28     671  1.7     3910  9.9         9  |
|    18   -.3881    .3341|  376   28     673  1.7     4583 11.6        11  |
|    19   -.2770    .3329|  385   27     720  1.8     5303 13.4        12  |
|    20   -.1663    .3325|  394   27     829  2.1     6132 15.5        14  |
```

Figure 4–67. Sample WINSTEPS Scoring Table

Pre-Equating Checks using Early Return Samples

With pre-equating, all items must have known calibrations that were established from their previous uses. The items are assumed to not interact differentially with instruction or changes in the environment over the intervening time period, and the underlying construct is assumed to not have changed due to changes in the content standards or mixes of item content or types. These are relatively strong assumptions that when met allow the raw-to-scale score conversion tables to be computed (as described above) as soon as the forms are constructed. This will permit preparing student reports as soon as scoring is complete and results have been certified.

Although pre-equating is possible when items have been previously administered and calibrated, a number of factors can affect the results. Some of these include changes in instructional practice (appropriate or not), news events, changes in popular culture, and disclosure of items. When these effects may be present and the pre-equating assumptions might not hold, it is important to verify the integrity of the pre-equating results.

To verify pre-equating, DRC recommends an early return sample be used to support post-equating analyses, which can then be compared to those from the pre-equating.

The *post-equating* procedure ensures scale scores and performance standards have the same meaning and interpretation in each administration. This process uses embedded field testing to generate item difficulty parameters for new items that could be used to construct new forms with minimal overlap with prior operational versions. After the spring administration, the item difficulty parameters can be recomputed and examined for consistency in the new context. If necessary, adjustments can be made to ensure the performance standards have the same meaning from year to year. However, DRC expects that the pre-equating will be validated by this process and will be used for reporting.

DRC recommends evaluating item stability, raw-to-scale score conversions, and the change in the percentage of students in each performance group. DRC has proposed the inclusion of designated post-equating anchor items in the core (common) section to allow for this check.

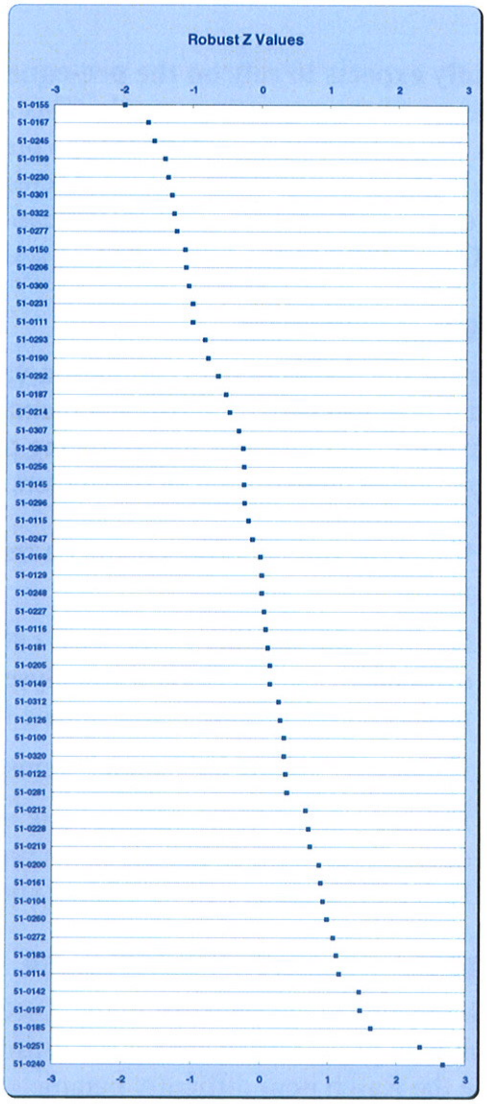


Figure 4–68. Robust Z Dot Plot

To evaluate the stability of the item parameters, DRC recommends examining the pre-equating Robust Z values for each item. A "dot" plot, similar to the example provided above in Figure 4–68, might be used to quickly visualize these results. Items are identified by their ID numbers and rank ordered according to their Robust Z values so that extreme values are more easily located. Historically, DRC has considered Robust Z values greater than 1.645 in absolute value to be unstable enough to warrant further evaluation (e.g., consultation with content experts on staff).

To further evaluate the appropriateness of the pre-equating results, **DRC recommends comparing the raw-to-scale score conversions from the pre-equating and post-equating procedure.** A sample graphical representation is presented in Figure 4–69, provided in the subheading below regarding the viability of pre-equating. Grid lines can be placed at scale scores representing performance level cuts, as illustrated in the example.

DRC fully expects to rely on the pre-equating process to maintain the consistency of the reporting metric. This acknowledges the possibility that there may be minor discrepancies between the pre-equating results and post hoc checks. In the unlikely event that the discrepancies would have a significant impact on the performance level classifications, DRC will consult with NDE and the TAC to determine the best course of action.

Item-Bank Maintenance

Monitoring and updating item calibration values to adjust for issues such as item parameter drift can help establish and maintain a successful pre-equating program. DRC recommends the following procedures be used to ensure that the most appropriate item difficulty parameters are "banked" for later use:

- Using the full data file, conduct a free local calibration for all operational items.
- Evaluate the stability of the local calibration results vs. the "banked" difficulties using a Robust Z analyses. (The current procedure, under the post-equating design, is to evaluate the stability of the anchor items.)
- Using only values for "stable" items, determine the *mean shift*.
- For operational items, update bank values by applying the mean shift to all operational items to put the items on bank scale. The resulting transformed Rasch difficulties will be banked and applied in future applications.

Calibrating Embedded Field Test Items

To calibrate embedded field test items, DRC recommends the use of a fixed common item procedure to anchor the estimates for all operational items and then estimate the Rasch item difficulty parameters for the field test items. In a pre-equating design, this entails anchoring all operational items to their verified banked values. This approach puts all field test items on the operational scale.

Establishing the Viability of Pre-Equating Design for the NeSA

DRC has conducted retrospective analyses comparing pre- and post-equating results on an informal basis for a sample of tests, similar to the NeSA design. **The results illustrate that the raw score cut points for each performance level from the pre-equating were equivalent to those established from the post-equating conducted after the regular administration.**

A graph from one of DRC's large-scale assessment clients comparing the raw-to-scale score conversions from the pre-equating and post-equating procedures for one mathematics test is shown in Figure 4–69. The overlapping curves indicate both equating methods yielded similar results. The C1, C2, and C3 lines mark the scale score cut points for each performance level. With the same raw score cut points, the percentages of students in each of the performance levels would also be equivalent across both equating methods.

While these retrospective studies would not satisfy all the requirements and assumptions for every pre-equating design, they do provide very strong evidence for the viability of the proposed pre-equating design for the NeSA.

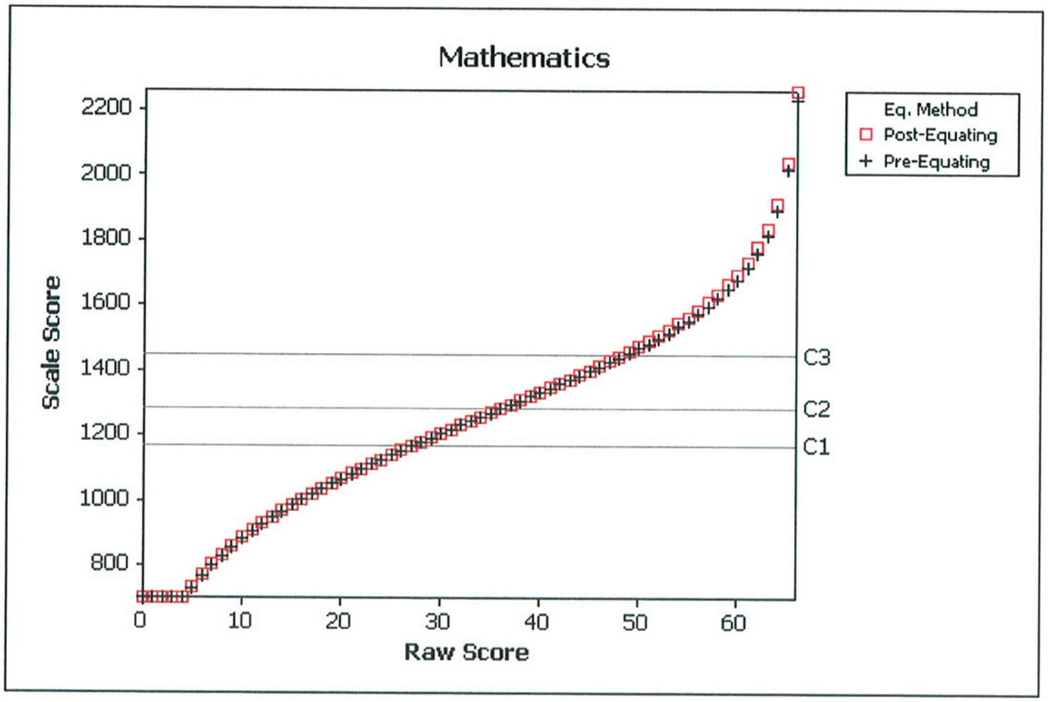


Figure 4–69. Sample Comparison of Raw-to-Scale Score Conversions from the Pre-Equating and Post-Equating Procedures

Because of the importance of the NeSA, it would be prudent to conduct a **formal and complete retrospective analysis** to establish the validity of the proposed pre-equating design.

c. Item Evaluation

i. Field Test Item Analysis

DRC Psychometric staff, led by **Dr. Ronald Mead,** will be providing all needed analysis of field test items for the NeSA and the accommodated assessments. DRC Psychometrics staff will work with DRC's Test Development staff to coordinate item analysis and forms construction. Statistical and psychometric analyses go through many phases in a testing program. DRC will analyze all items prior to being placed on forms using the methodologies described in Table 4–8.

Table 4–8. Item Analyses

Classic Item Analyses (Overall and by Subgroup where Requested)	
p-values, with flags for very easy and very difficult items	Percent choosing each multiple-choice (MC) option, with flags for distractor percent higher than correct-answer percent
Corrected item-total correlations, with flags for possible mis-key or poor item quality (point-biserial)	Option-total correlations for MC items
Test reliability	Standard error of measurement for the scale
Differential Item Functioning (for Field Test Items)	
Focal group designation	Reference group designation
Favored group designation	Mantel-Haenszel chi-square for MC items
Mantel-Haenszel delta for MC items	ETS DIF category (A, B, C) for MC items
Rasch Statistics	
Item Statistics:	
Logit difficulty estimates	INFIT and OUTFIT statistics
Standard errors for all parameter estimates	
Test Indices:	
Test information function	Test characteristic curves
Raw-Logit-scale score tables	Standard errors for all parameter estimates and scale scores
Person separation reliability	

Classic Item Analyses

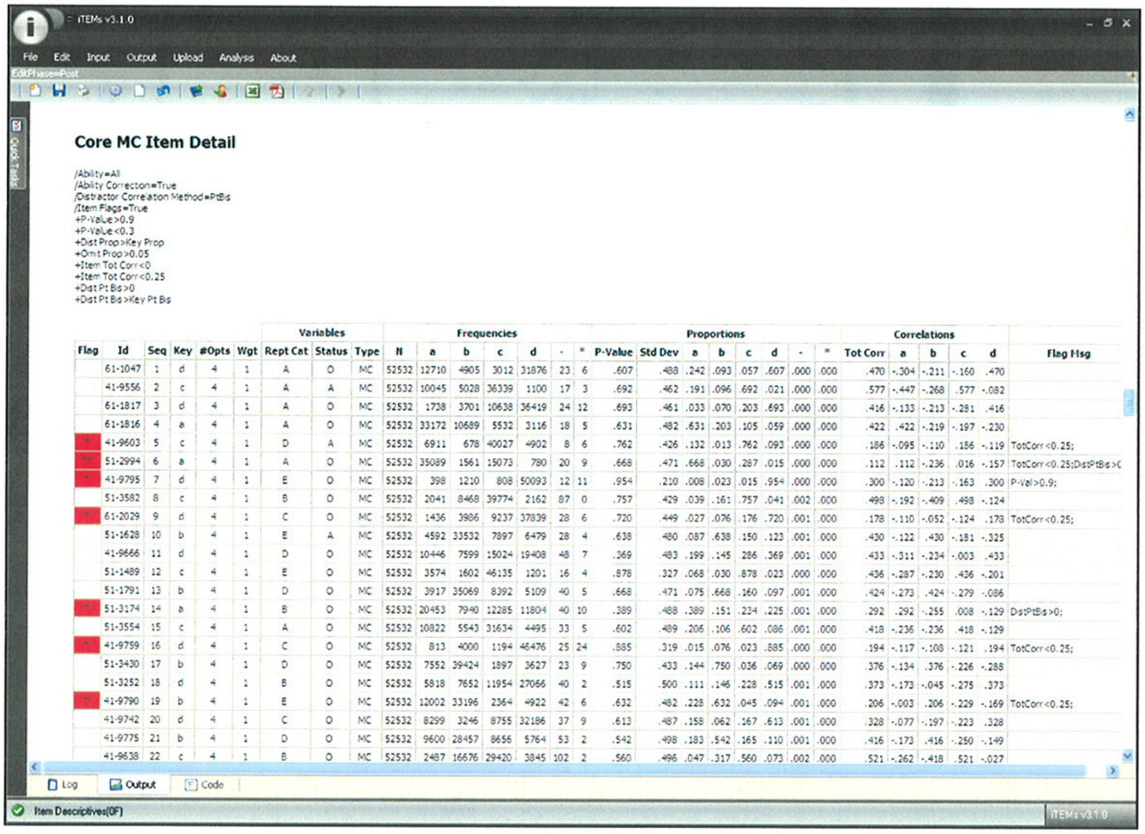
Embedded field test item analyses will be conducted by form to ensure problems in one form are not masked by other forms. This begins with classic item analysis. DRC may use its proprietary Item and Test Evaluation Modules (iTEMs) system. DRC's Psychometric staff can start with the key verification module of this system, computing the number and proportion of students selecting each response option, the *p*-value for the item, the item-total correlation for the key, and the item-total correlations for each of the response alternatives. These statistics can be used to flag any potential incorrect scoring keys. Typically DRC flags items as possibly mis-keyed if the following conditions are observed:

- Percent correct (*p*-value) is low;
- Percent of students selecting any distractor is high;
- Point-biserial correlation for the key is low or negative;
- Point-biserial correlation for a distractor is high.

With iTEMs, the criteria for flagging an item are customizable. As an example, the "low" *p*-value threshold could be set at any value (e.g., 0.30, 0.35, 0.40). DRC psychometricians will work with NDE to define the criteria that are most suited for the NeSA.

Key Verification

Figure 4–70 presents a copy of an on-screen display of the iTEMs key verification module. The red flag designates an item flagged that met one or more of the above mentioned criteria.



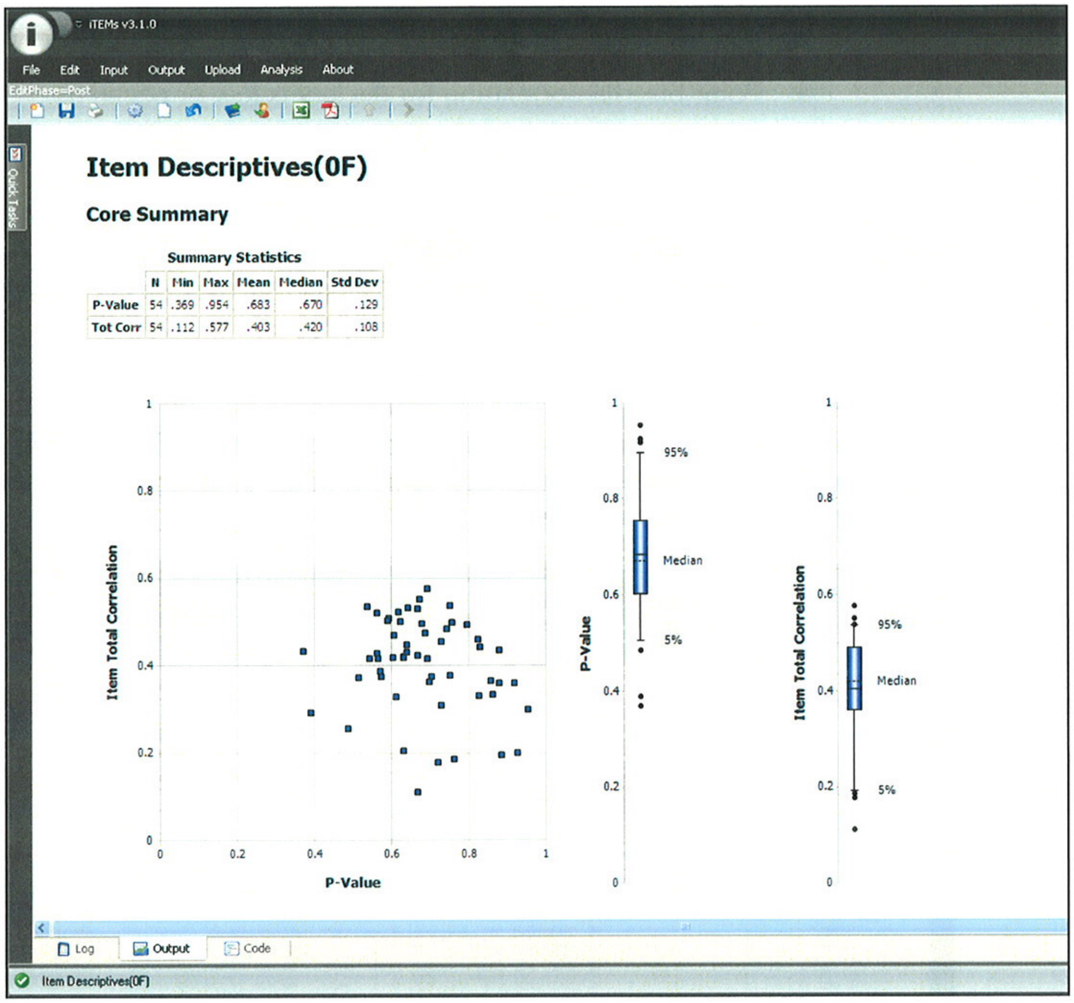
Core MC Item Detail

/Ability=All
/Ability Correction=True
/Distractor Correlation Method=PtBis
/Item Flags=True
+P-Value>0.9
+P-Value<0.3
+Dist Prop>Key Prop
+Omit Prop>0.05
+Item Tot Corr<0
+Item Tot Corr<0.25
+Dist Pt Bis>0
+Dist Pt Bis>Key Pt Bis

						Variables			Frequencies							Proportions								Correlations					
Flag	Id	Seq	Key	#Opts	Wgt	Rept Cat	Status	Type	N	a	b	c	d	-	*	P-Value	Std Dev	a	b	c	d	-	*	Tot Corr	a	b	c	d	Flag Msg
	61-1047	1	d	4	1	A	O	MC	52532	12710	4905	3012	31876	23	6	.607	.488	.242	.093	.057	.607	.000	.000	.470	-.304	-.211	-.160	.470	
	41-9556	2	c	4	1	A	A	MC	52532	10045	5028	36339	1100	17	3	.692	.462	.191	.096	.692	.021	.000	.000	.577	-.447	-.268	.577	-.082	
	61-1817	3	d	4	1	A	O	MC	52532	1738	3701	10638	36419	24	12	.693	.461	.033	.070	.203	.693	.000	.000	.416	-.133	-.213	-.281	.416	
	61-1816	4	a	4	1	A	O	MC	52532	33172	10689	5532	3116	18	5	.631	.482	.631	.203	.105	.059	.000	.000	.422	.422	-.219	-.197	-.230	
*	41-9603	5	c	4	1	D	A	MC	52532	6911	678	40027	4902	8	6	.762	.426	.132	.013	.762	.093	.000	.000	.186	-.095	-.110	.186	-.119	TotCorr<0.25;
**	51-2994	6	a	4	1	A	O	MC	52532	35089	1561	15073	780	20	9	.668	.471	.668	.030	.287	.015	.000	.000	.112	.112	-.236	.016	-.157	TotCorr<0.25;DistPtBis>(
*	41-9795	7	d	4	1	E	O	MC	52532	398	1210	808	50093	12	11	.954	.210	.008	.023	.015	.954	.000	.000	.300	-.120	-.213	-.163	.300	P-Val>0.9;
	51-3582	8	c	4	1	B	O	MC	52532	2041	8468	39774	2162	87	0	.757	.429	.039	.161	.757	.041	.002	.000	.498	-.192	-.409	.498	-.124	
*	61-2029	9	d	4	1	C	O	MC	52532	1436	3986	9237	37839	28	6	.720	.449	.027	.076	.176	.720	.001	.000	.178	-.110	-.052	-.124	.178	TotCorr<0.25;
	51-1628	10	b	4	1	E	A	MC	52532	4592	33532	7897	6479	28	4	.638	.480	.087	.638	.150	.123	.001	.000	.430	-.122	.430	-.181	-.325	
	41-9666	11	d	4	1	D	O	MC	52532	10446	7599	15024	19408	48	7	.369	.483	.199	.145	.286	.369	.001	.000	.433	-.311	-.234	-.003	.433	
	51-1489	12	c	4	1	E	O	MC	52532	3574	1602	46135	1201	16	4	.878	.327	.068	.030	.878	.023	.000	.000	.436	-.287	-.230	.436	-.201	
	51-1791	13	b	4	1	D	O	MC	52532	3917	35069	8392	5109	40	5	.668	.471	.075	.668	.160	.097	.001	.000	.424	-.273	.424	-.279	-.086	
*	51-3174	14	a	4	1	B	O	MC	52532	20453	7940	12285	11804	40	10	.389	.488	.389	.151	.234	.225	.001	.000	.292	.292	-.255	.008	-.129	DistPtBis>0;
	51-3554	15	c	4	1	A	O	MC	52532	10822	5543	31634	4495	33	5	.602	.489	.206	.106	.602	.086	.001	.000	.418	-.236	-.236	.418	-.129	
*	41-9759	16	d	4	1	C	O	MC	52532	813	4000	1194	46476	25	24	.885	.319	.015	.076	.023	.885	.000	.000	.194	-.117	-.108	-.121	.194	TotCorr<0.25;
	51-3430	17	b	4	1	D	O	MC	52532	7552	39424	1897	3627	23	9	.750	.433	.144	.750	.036	.069	.000	.000	.376	-.134	.376	-.226	-.288	
	51-3252	18	d	4	1	B	O	MC	52532	5818	7652	11954	27066	40	2	.515	.500	.111	.146	.228	.515	.001	.000	.373	-.173	-.045	-.275	.373	
*	41-9790	19	b	4	1	E	O	MC	52532	12002	33196	2364	4922	42	6	.632	.482	.228	.632	.045	.094	.001	.000	.206	-.003	.206	-.229	-.169	TotCorr<0.25;
	41-9742	20	d	4	1	C	O	MC	52532	8299	3246	8755	32186	37	9	.613	.487	.158	.062	.167	.613	.001	.000	.328	-.077	-.197	-.223	.328	
	41-9775	21	b	4	1	D	O	MC	52532	9600	28457	8656	5764	53	2	.542	.498	.183	.542	.165	.110	.001	.000	.416	-.173	.416	-.250	-.149	
	41-9638	22	c	4	1	B	O	MC	52532	2487	16676	29420	3845	102	2	.560	.496	.047	.317	.560	.073	.002	.000	.521	-.262	-.418	.521	-.027	

Figure 4–70. iTEMs Key Verification Module

Figure 4–71 shows a display of the iTEMs classical analysis module. It presents the *p*-values and item-total correlations of items allowing for the visual detection of outliers and any other unexpected relationship.



Item Descriptives(0F)

Core Summary

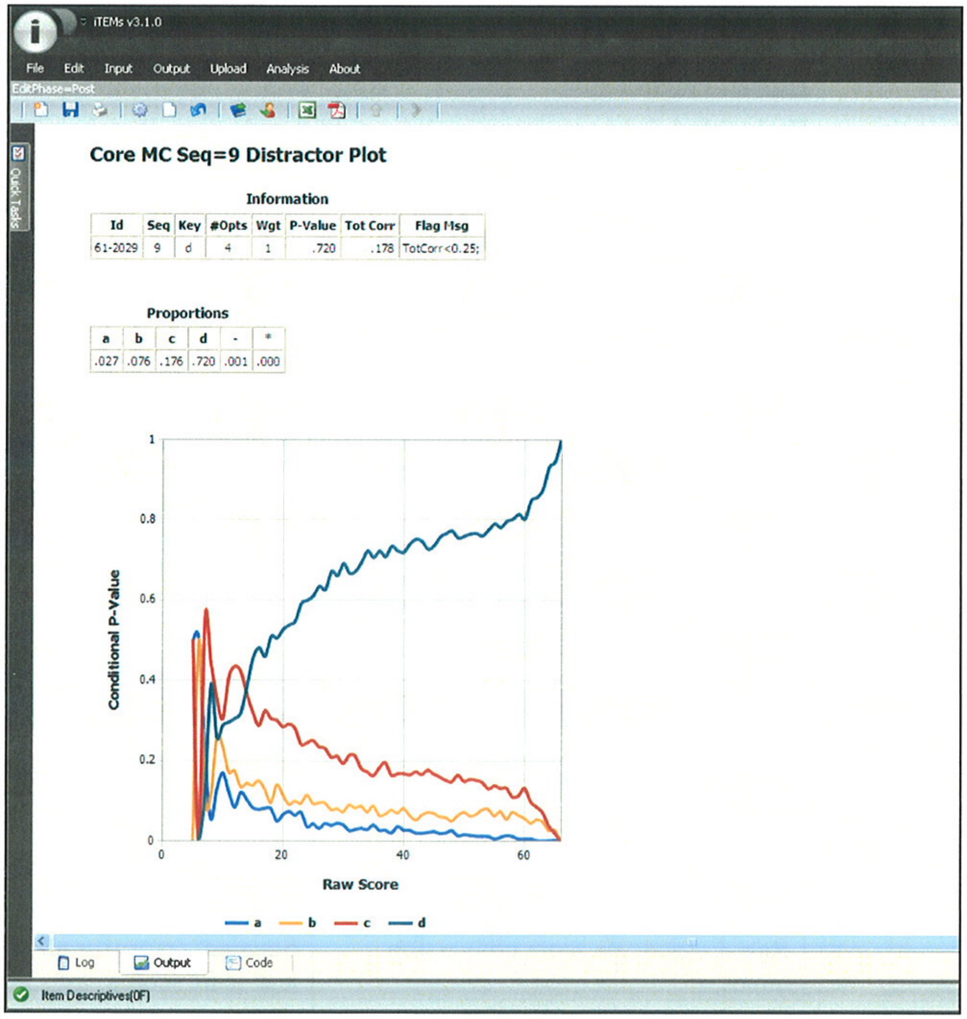
Summary Statistics

	N	Min	Max	Mean	Median	Std Dev
P-Value	54	.369	.954	.683	.670	.129
Tot Corr	54	.112	.577	.403	.420	.108

Figure 4–71. iTEMs Classical Analysis Module

Distractor Analysis

In the distractor analysis shown in Figure 4–72, iTEMs generates a graph depicting the proportion of students selecting each response option as a function of raw score. The proportion of students selecting the keyed response option should increase as a function of ability (raw score) increases. Conversely, the proportion of students selecting each of the incorrect response options (distractors) should decrease as ability increases. A graph for an item that does not show this pattern of results may indicate an incorrect key. DRC has found that these item distractor analysis graphs, when used in conjunction with the above-mentioned item statistics, are a powerful tool in detecting possible item mis-keys.



Core MC Seq=9 Distractor Plot

Information

Id	Seq	Key	#Opts	Wgt	P-Value	Tot Corr	Flag Msg
61-2029	9	d	4	1	.720	.178	TotCorr<0.25;

Proportions

a	b	c	d	-	*
.027	.076	.176	.720	.001	.000

Figure 4–72. iTEMs Distractor Analysis

The item analysis will be conducted as soon as data based on an appropriate calibration sample are available. This analysis will be conducted by form. All items flagged as possibly mis-keyed are immediately referred to DRC Test Development content specialists, Project Management, Information Systems, and Software Quality Assurance staff for further review and verification. Incorrect item keys are identified and evaluated before the final scoring is conducted. Therefore, there are no implications for item calibrations, scaling, equating, and reporting. Documentation related to any item discrepancies and a copy of the item analysis will be available to NDE for review upon request.

iTEMs can provide *p*-values, distractor analyses, item-test correlations, percent of students at each constructed response score point (if used), the standard error of measurement, the alpha coefficient, and Differential Item Functioning statistics (DIF). DIF analysis is discussed in greater detail in *Subheading 2.d., Content of Test Forms.* In addition, analyses by subgroup can be conducted by gender, ethnicity, ELL status, IEP status, economic disadvantage, and/or other subgroups as requested by NDE.

Rasch Analysis

Embedded field test items will be calibrated using Rasch methodology. For a detailed description of Rasch calibrating, please see *Subheading 7.a.*

ii. Operational Analysis

Item analysis of operational items on the NeSA, and the accommodated versions, is shown in Table 4–9. Analyses will include the following:

Table 4–9. Item Analyses

Classic Item Analyses (Overall and by Subgroup where Requested)	
p-values, with flags for very easy and very difficult items	Percent choosing each multiple-choice (MC) option, with flags for distractor percent higher than correct-answer percent
Corrected item-total correlations, with flags for possible mis-key or poor item quality (point-biserials)	Option-total correlations for MC items
Test reliability	Standard error of measurement for the scale
Rasch Statistics	
Item Statistics:	
Logit difficulty estimates	INFIT and OUTFIT statistics
Standard errors for all parameter estimates	
Test Indices:	
Test information function	Test characteristic curves
Raw-Logit-scale score conversion tables	Standard errors for all parameter estimates and scale scores
Person separation reliability	

Classic Item Analyses

As discussed above in *Subheading 7.c.i.*, DRC may use its proprietary Item and Test Evaluation Modules (*iTEMs*) for classic item analyses of operational items. The key verification module of this system computes the number and proportion of students selecting each response option, the *p*-value for the item, the item-total correlation (e.g., the point-biserial correlation) for the key, and the item-total correlations for each of the other response alternatives. These statistics are used to flag any potentially incorrect scoring keys. DRC psychometricians will work with NDE to define the criteria that are most suited for the NeSA. Also discussed previously, *iTEMs* can customize the exact criteria for flagging an item.

In the *distractor* analysis, *iTEMs* generates a graph depicting the proportion of students selecting each response option as a function of raw score. The proportion of students selecting the keyed response option should increase as a function of ability. Conversely, the proportion of students selecting each of the distractors should decrease as a function of ability. A graph for an item that does not show this pattern of results may indicate an incorrect key. DRC has found that these item distractor analysis graphs, in conjunction with the traditional item statistics, are powerful tools in detecting possible item mis-keys.

The item analysis will be conducted as soon as data based on a sufficient calibration sample is available. All items flagged as possibly mis-keyed will be referred to DRC content specialists, Project Management, Information Systems, and Software Quality Assurance staff for further review and verification. Possible incorrect item keys will be identified, confirmed, and corrected before the final scoring is conducted. Therefore, there will be no implications for item calibrations, scaling, equating, and reporting. Documentation related to any item discrepancies and a copy of the item analysis will be available to NDE for review upon request.

iTEMs can provide *p*-values, distractor analyses, item-test correlations, percent of students at each constructed response score point (if used), the standard error of measurement, and the alpha coefficient all used in a post-equating check.

Rasch Statistics

Rasch statistics will be calculated using *WINSTEPS*, a software package commonly used in the industry. For all items, threshold difficulty parameters will be provided with their associated standard errors of estimation. In addition, *infit* and *outfit* statistics will be provided. DRC will flag items for very low or very high difficulty estimates or for unexpectedly extreme threshold parameter estimates. For more on Rasch analysis, please see *Subheading 7.a.,* above.

d. Test Construction

DRC's Psychometric Services staff play an integral role in the test construction process. *Subheading 2.d., Content of Test Forms,* provides a detailed discussion of our proposed test construction process.

e. Scoring

Psychometric Quality and Methodology

DRC's Psychometric Services (PS) Department is committed to quality and excellence. We achieve psychometric quality and excellence by assuring that our practices and procedures meet the professional measurement standards outlined in the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999). **In addition, in 2006, DRC took the initiative and led the industry by starting a Department of Psychometric Quality.** The department is headed by **Dr. Adisack Nhouyvanisvong, Director of Psychometric Quality,** who oversees all aspects of psychometric quality for DRC. Dr. Nhouyvanisvong has provided technical training and support for numerous large-scale testing programs. With more than eight years of experience in psychometrics, Dr. Nhouyvanisvong has expertise in item calibration and equating, classical and IRT item analyses, DIF analysis, standard setting, computer-based testing (CBT) and computer-adaptive testing (CAT). In addition to his assessment and measurement work in the testing industry, Dr. Nhouyvanisvong has experience teaching college-level and graduate-level courses in research methods and advanced measurement theory and applications.

DRC maintains a department devoted solely to ensuring psychometric quality in all of our large-scale assessment programs.

At DRC, Dr. Nhouyvanisvong's primary focus is the continual improvement of psychometric quality processes, working closely with DRC's psychometric team. Dr. Nhouyvanisvong currently oversees and ensures psychometric integrity for numerous statewide assessments, including programs for Idaho, Louisiana, Oklahoma, Pennsylvania, South Carolina and Washington. One of Dr. Nhouyvanisvong's main initiatives has been the creation of a data forensics system at DRC. This system, called **Psychometric Scenes Investigator (PSI) will systematically and efficiently conduct numerous analyses to ascertain the integrity of test results.** A more detailed discussion of PSI can be found in our proposed option for data forensics, included in *Appendix F.* If NDE is interested in this option, costs can be provided upon contract award.

The Psychometric Services Quality (PSQ) staff consists of research analysts skilled in research and measurement theory and methodology. Working closely with our psychometricians, they will provide psychometric quality control for all the assessments covered in this contract through the following procedures, described in Figure 4–73.

- **Data files**—Quality checks will be performed by this team to verify the integrity of data files.
- **Scored data**—Quality checks will be performed on the data to ensure that test scores have been scored correctly.
- **Calibration, Scaling, and Equating**—Replication of these processes will be performed as an independent quality check.
- **Independent Psychometric Analysis**—If needed, coordinate with the third-party vendors performing independent psychometric analyses.
- **Reports**—Validate that the assessment results are accurate and allow for valid interpretations.
- **Item Statistics**—Ensure item statistics are properly stored in the item bank system.
- **Trend Analysis**—Quality checks will be performed on the data and scoring to ensure no anomalies exist relative to historical performance trends.

Figure 4–73. Psychometrics Quality Procedures

Data File Quality Control

Psychometric quality begins with a check of the student response data file. All fields critical to the analysis, calibration, and equating process are checked and verified by the psychometric and the psychometric quality teams. Variables are validated against the final approved file layout and processing rules to ensure that no unanticipated values exist and that data characteristics appear to be consistent with past experience. All key demographic fields are checked for accuracy. Additionally, a general reasonableness check on the data is conducted by computing the raw score frequency distribution, verifying the proper numbers of items and the proper location of open-ended items, and by verifying that no unusual or atypical values exist.

Preliminary Item Analysis for Key Verification

Psychometric quality control continues with a preliminary item analysis key check on multiple-choice items. There are many levels of key verification that take place within DRC (i.e., content experts take the exam and compare their keys with the approved scoring key and SQA staff tests and verifies the scoring program), but this preliminary item analysis serves as a final check to identify any items that do not seem to be functioning as expected. The preliminary item analysis is performed on a scored student file as soon as enough student records are available. Items that exceed certain criteria in terms of psychometric characteristics are flagged and re-verified by content experts to ensure that the identified item key is correct.

The preliminary item analysis is conducted by our psychometric team and verified by our psychometric quality team. The process is an integral part of ensuring quality and the validity of the test results. The analysis provides assurances that the test is of high quality, and therefore valid inferences can be made from the results.

Quality Check on Post and Pre-Equating

For information on quality checks for pre-equating, please see *Subheading 7.b., Equating*.

Student Data Files and Reporting Quality

The students' raw scores (number correct) will be converted to scale scores derived by a linear transformation of the Rasch logit abilities. Operationally, the scale scores will be placed in raw-to-scale score conversion tables and the appropriate values retrieved and posted to each student record. PSQ and SQA analysts will be involved in verifying the tables and the look-up process. The scale scores will be applied to the final student data files, individual student reports, and summary reports. All files and reports go through multiple levels of quality checks. The PSQ research analysts serve as one part of that process by performing independent checks on the data files and reports.

Item Bank Process

As previously discussed, all calibrations will be run independently by the psychometrician and statistical analyst and subsequently verified by the research analysts. After these independent runs are conducted, the entire team will review and evaluate the results. This process will be directed by the Psychometric Lead and the Director of Quality for Psychometric Services. Any discrepancies will be noted, discussed, and resolved before the calibrations will be considered final and imported into the item bank system.

Once statistics are put into the item bank, all data elements will be checked to ensure that they have been imported without error. Lastly, sample data cards will be printed from the system and checked to ensure that proper statistics and values are displayed correctly.

f. Reporting

a. Analyses to Support Reporting Results

DRC's Psychometric Services and Information Systems teams work hand-in-hand during the analysis and reporting phase of administration. No reports are released without all analyses described earlier in this section having been performed and signed off by the Lead Psychometrician.

NeSA Comparability Studies

Whenever tests that are administered under both testing modes (computer-based and traditional paper and pencil) co-exist in an assessment program, score comparability between computerized and paper-and-pencil tests becomes an important issue. Under the dual-mode testing environments, scores from the two modes cannot be used interchangeably for interpretation and/or reporting purposes without supportive evidences from carefully designed and conducted empirical research over the target testing population (AERA, APA, NCME, 1999). CAL staff has extensive experience and expertise in conducting paper-pencil (P&P) and online computer delivery (CBT) mode comparison studies within the context of high-stakes state testing environments. Descriptions of various comparability designs (i.e., double testing or test-retest, matched groups, volunteer groups, and randomly assigned groups) are presented below, followed by the proposed comparability study for the NeSA administrations.

CAL staff has conducted paper-pencil (P&P) and online computer delivery (CBT) mode comparison studies yearly over the life of the online testing program in the state of Kansas. The design and results from the studies during the first year (2003) for the grade 7 mathematics test can be found in Poggio, Glasnapp, Yang and Poggio (2005) at http://www.jtla.org. The design and results from the studies during the second year (2004) for tests at three grade levels of both reading and mathematics tests can be found in Poggio, Glasnapp, Yang, Beauchamp and Dunham (2005). When these studies were conducted, the only viable design for data collection was to “double test” students on parallel forms of the test, once under a P&P format and once under the CBT format. District/school participation in the studies was voluntary. Although results lacked statistical difference based on very large sample sizes (numbering in the tens of thousands), a limitation of these two studies was that order of testing mode was not controlled nor was the order data reported by schools viewed to be trustworthy.

The double tested design and data collection were replicated again during the 2005 testing period for tests at three grade levels in reading and mathematics, but data also were collected for test forms at three grade levels in science and social studies. Information to study and control the order effect was captured in this series of studies. In this work, again we did not observe statistically significant results between the computerized testing and paper and pencil modes for grades or content areas.

In the Kansas comparability studies, the conditions and constraints of the testing program when the studies were initiated necessitated that a “double testing” design be put in place so that data would exist such that the individual students served as their own control in the repeated measures design. In this design, order of administration is a potential problem and would best be controlled through random assignment. Because of the administration time involved in double testing and the uncertainty of having adequate controls for the order effect through random assignment of order, more efficient data collection designs are available if the right conditions exist in a state's testing program.

An alternative and attractive quasi-experimental design does exist if the right conditions can be put in place. If students' prior years' test scores are available, these scores can be used as matching control variables or covariates to control for potential prior achievement differences in the volunteer CBT group and the selected P&P comparison group along with other matching covariates (propensity scores for matching and controlling). While the immediate prior year test scores in Nebraska would not be available in the first year of respective operational NeSA assessments, prior scores from earlier assessments likely would be available and could be used as matching control variables along with select demographic variables to control for achievement differences in the groups taking tests under the different mode conditions. This approach would have considerable merit and value during early years of operational implementation.

Another study possibility is to use the data as it exists from volunteer groups knowing that they likely represent non-equivalent groups, but attempt to demonstrate comparability by looking at the structural consistency of the tests across modes, (i.e., conduct the studies addressing structural validity with test mode as a variable). If differences are not found, evidence is provided to support a conclusion that the tests are functioning similarly. Such approaches have been taken in the Kansas studies (2006) and have demonstrated an exceptionally high degree of structural consistency of the test items across modes.

There is no doubt that the best, failsafe design is to implement a true randomized experimental design with random selection and assignment of student to test mode. It is the design of choice and the one we would recommend is implemented if at all possible. However, random assignment of administration mode (paper and pencil or online) is preferable at the student level, but often is not practical. If this design is desired by NDE, we will work with NDE to design the most feasible random sampling and assignment design (at the student, class, or building level) to be implemented during test administrations. The assigned groups would be of sufficient size and representativeness as to be considered randomly equivalent. We would then explore the structural similarity of the constructs being assessed by the NeSA assessments across delivery modes through the use of appropriate factoring techniques, similar to the procedures employed in the volunteer groups design described above.

In addition to the work by CAL personnel, it should also be noted that DRC was awarded a contract from the state of South Carolina (2006–07) to explore the feasibility of moving P&P testing to the online CBT delivery mode. As part of that contract, an exhaustive literature review was made examining studies addressing the comparability issue. This latter review will serve as a foundation to provide CAL, DRC, and NDE with information in deciding which design to best implement within the context of the Nebraska online assessment implementation schedule.

This presentation on comparability is intended to convey the vast experience and expertise of CAL and DRC staff in addressing the comparability issue. DRC and CAL will work with NDE to implement the most efficient and valid design within the context of the Nebraska testing program during the contract period to address the comparability of paper and pencil and online test delivery mode. In the end, we would plan to implement a design that NDE and its advisors wholeheartedly support and endorse.

8. Reporting

DRC has 30 years of experience in reporting large-scale assessment results. Our experience on other assessments, such as those for Alaska, Idaho, Louisiana, Oklahoma, Pennsylvania, and South Carolina, can assure NDE that DRC has the ability to report accurate results in critically prescribed time limits. Our comprehensive reporting package is a collaboratively crafted system that offers **flexibility**. The reports' design and content will be **useful and easy to understand** and will be **produced and delivered to each district/school on time.** We have provided sample reports in *Appendix G* of our proposal.

DRC state assessment clients appreciate our ability to tailor reporting solutions to meet their needs, while still maintaining superior quality and timely delivery.

For each new project, DRC works with our clients to **customize our reporting process to the unique needs of their assessments**. We offer the combination of proven excellence in designing and implementing customized solutions to meet expectations, in-depth understanding of the complexities of assessment reporting, and a cadre of highly qualified professionals who are experienced and will work collaboratively with NDE to address all reporting requirements, as well as the needs of students, parents, and educators.

At the core of our proposal is our commitment to continue to provide NDE with **an innovative, customized reporting package**. Highlights of our reporting package include the following:

- **Customized Reports** that are user-friendly, delivered on time, and able to meet the evolving needs of NDE. They will be fully aligned with Nebraska's reporting categories and will include easily understood data presentations.
- **A Report Process** that is established, efficient, and provides high-quality reports.
- **Web-Based Report Delivery System** that provides for the timely release of results.
- **Data Analysis Tools,** offered by CAL, that provide options for schools and districts to analyze their results.
- **Test Interpretation Manuals and Item Samplers,** offered online as PDFs for NDE and schools/districts.

Reporting Quality Procedures

DRC incorporates rigorous quality assurance activities throughout the reporting process to ensure the highest level of quality and data integrity. The focus on "building in quality" and "issue prevention" ensures our clients quality products and services.

Our primary goal is to ensure the quality of student data and to make certain that each student record is tested and verified for completeness and accuracy. DRC's familiarity with reporting requirements and data elements similar to those required by NDE provide our Software Quality Assurance Analysts with a solid platform and experience that will be invaluable to the NeSA program. Upon the completion of the thorough data verification process, quality checks will be performed on the data placement and report file formatting for each data element displayed on the reports. All reporting data elements will be verified back to the production data file and the reporting processing rules. Additional quality cross-checks will be performed to ensure accuracy and consistency across all reporting mediums for the assessment. This includes posting data to our secure web-based Report Delivery System, hardcopy reports, or any other type of reporting medium.

Similar quality checks will also be used to validate data at the school, district, and state level. Senior Software Quality Assurance Analysts will conduct a second review of each report to ensure methodology, processes, and procedures are followed and verify that the reports are approved for production. An additional post-print review is conducted before any hardcopy reports are packaged and shipped.

Report Generation Process

Report Design

The varied audiences receiving assessment data have differing needs and requirements when viewing or analyzing the data. DRC currently produces a wide array of reports of varying types including rosters, summaries, and disaggregations at various levels, including state, district, school, classroom, and individual student. State, district, school, and classroom reports typically contain aggregated or summary information, including averages and distributions of percent correct answers, scale scores, performance levels, etc. Individual student reports typically contain specifics of the individual student's performance as well as a comparison to the school, district, and state average performance. DRC plans to utilize a standard black-and-white report design for the NeSA, based on information provided by NDE in the RFP and Q&A. However, if desired by NDE, we could also provide a full color design for parent/guardian or summary reports. Samples of color student/parent reports that we provide in Louisiana and Oklahoma are provided in *Appendix G.*

DRC will leverage our many years of experience working with other state departments to provide suggestions on the content, layout, and appearance of reports in order to maximize benefit and value.

Reporting Requirements

DRC's staff has a wealth of experience in defining and documenting requirements for data analysis and report development. DRC will work closely with NDE on these requirements and produce a document that explicitly describes all of the processing rules used for the design and development of the scoring and reporting software. This document will be used as the standard for all software development, the definition of acceptance testing criteria, and the development of scripts for test plans during the internal quality assurance process. DRC's Psychometric Services staff will ensure that data flow from materials receipt through reporting complies with standards for educational and psychological testing. Definition of content and format of data files and hardcopy reports will also be developed and documented during this time.

Report Mockups

Report mockups are essential in the report development process. DRC will create report mockups representative of the exact production reports that will be delivered for each administration. The mockups will be comprised of simulated, but realistic, data elements. The mockups will be in the required report layout, display the appropriate fonts and font sizes, and demonstrate paper size and printing elements.

DRC will follow a process that provides NDE with the opportunity to review, change, and approve all mockups prior to report development. The mockups will be reviewed by DRC's System Analysts and Software Quality Assurance staff for accuracy, consistency, and ensure they are meeting the initial requirements. During the review process, NDE will be able to evaluate the static content and layout of each report to make certain it reflects the format, verbiage, and design required. DRC will work with NDE throughout the review process to incorporate any changes or modifications.

DRC's review process is outlined in Figure 4–74.

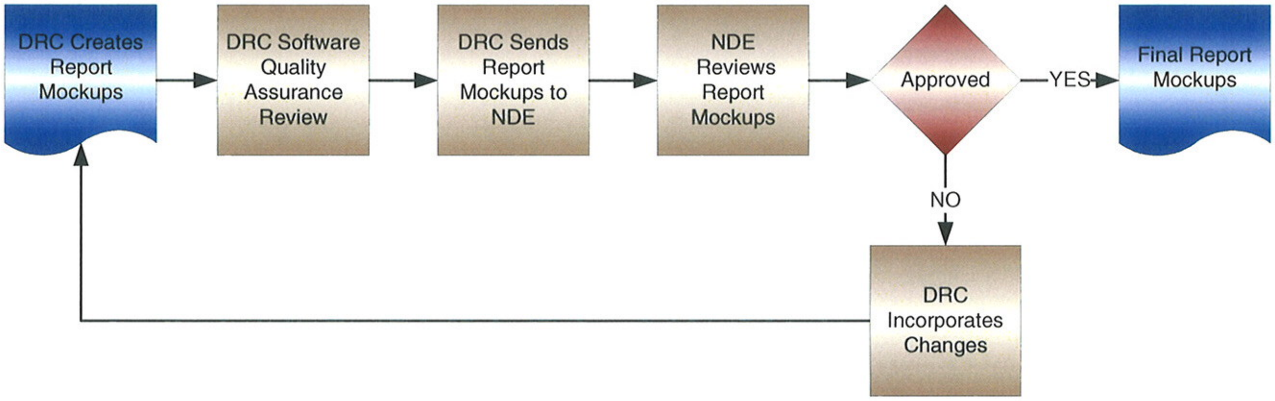

DRC Creates Report Mockups
DRC Software Quality Assurance Review
DRC Sends Report Mockups to NDE
NDE Reviews Report Mockups
Approved
YES
Final Report Mockups
NO
DRC Incorporates Changes

Figure 4–74. DRC's Report Review Process

Due dates for the report mockups will be clearly outlined in the schedule provided to NDE and negotiated among all appropriate parties. The report mockups will be completed, reviewed, and agreed upon by NDE to ensure the final reports meet the requirements.

Report Generation

DRC understands the activities and coordination required to accurately and comprehensively report large-scale assessment results. DRC has proven success with understanding and implementing reporting requirements and currently produces a wide variety of reports for clients, including individual student reports, labels, rosters at varying levels, summaries at varying levels, and item analysis reports. We offer to NDE our superior record of meeting reporting deadlines for large-scale, statewide assessments around the country.

DRC's staff has a wealth of experience in defining and documenting requirements for complex data analysis and report development and we have a solid knowledge base from which to build. We will work closely with NDE to refine the reporting requirements that explicitly describe all elements used for the design and development of the reporting software. The requirements document will be used as the standard for all software development, the definition of acceptance testing criteria, and the development of test scripts for the internal quality assurance process.

We employ a two-step report generation process. The first step is to perform all calculations and analysis to produce the data elements contained on the reports. The second step takes the data and formats it for presentation on the reports. This process allows the data to be thoroughly verified prior to and independent of formatting of the reports. It also allows for data calculations to be performed once, but yet presented in multiple formats.

Final Data and Report Review

The final data and reporting review with NDE is a critical component of our reporting process. DRC will perform a thorough quality assurance review prior to release of reports. All files and reports are thoroughly tested to guarantee accuracy.

Upon approval from NDE, DRC will produce the final student, class, school, district, and state reports. DRC's large-scale assessment reporting experience can ensure NDE that accurate and high-quality reports will be delivered within the prescribed time limits of the contract. Over the years, DRC has **repeatedly demonstrated the ability to provide ongoing communication and to deliver on time accurate data and reports** to states, districts, schools, and students/parents.

Figure 4–75 shows our report generation, review, and approval process.

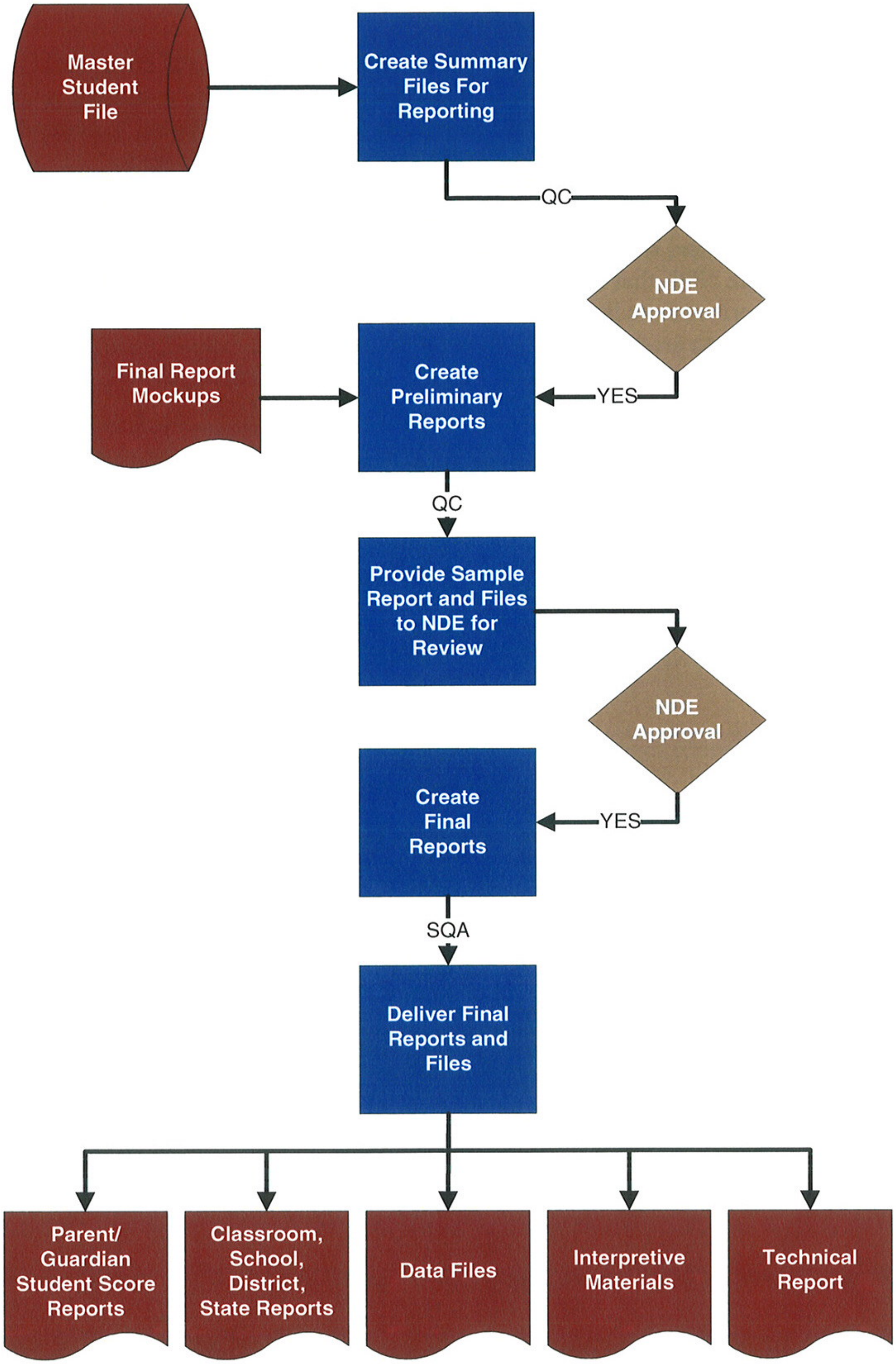

Master Student File
Create Summary Files For Reporting
QC
NDE Approval
Final Report Mockups
Create Preliminary Reports
YES
QC
Provide Sample Report and Files to NDE for Review
NDE Approval
Create Final Reports
YES
SQA
Deliver Final Reports and Files
Parent/ Guardian Student Score Reports
Classroom, School, District, State Reports
Data Files
Interpretive Materials
Technical Report

Figure 4–75. DRC Reporting Process

Reporting Timeline

DRC fully appreciates NDE's desire to get student results in the hands of key district and school personnel as quickly as possible following test administration. We understand that NDE requires that summary results be available to the State, as well as schools and districts in approximately 2 weeks following test administration, and parent/guardian reports delivered to districts by the end of the school year, based on the LB 1157 Work Plan. DRC is proposing an April test administration window to help accommodate this desire for rapid reporting.

However, it is important to note that there are several factors that will influence the report timeline during the initial years of implementation of the NeSA. First, student and summary results will not be reported for a given subject until after standards have been set following the first operational administration. Second, student and summary reports will contain all operational subjects (therefore, reporting will be delayed until standards have been set for all subjects). Given these factors, a two-week turnaround on reports cannot be accomplished until the 2012–2013 school year.

In order to overcome these issues, and to still get some level of detail in the hands of parents and district and school personnel in a timely fashion, DRC is proposing the creation of district student data files and online parent letters for the 2010–2011 and 2011–2012 school years. District student data files will contain results for subjects for which standards have been set (reading in 2010–2011 and reading and mathematics in 2011–2012). They will contain data, organized by school, regarding student demographics and performance. The online parent letters will also contain scores for subjects where standards have been set. They can be printed and distributed by schools and districts to students and their families as an advanced opportunity for seeing student results on the spring assessment. The parent letters will be available through DRC's online Report Delivery System. More information about our Report Delivery system can be found later in this section under *Subheading 8.a.x.*

Table 4–10 highlights DRC's proposed report deliverables and timelines by year. We have also included a more detailed milestone schedule in *Subheading 2.e.i.* DRC will be happy to discuss this plan with NDE and make any modifications upon contract award.

Table 4–10. Proposed Report Deliverables and Timeline

	Online Parent Letters and District Student Data Files	Online Summary Reports	Hardcopy Parent/Guardian Reports
Year 1 (2008–2009)	No reporting		
Year 2 (2009–2010)	n/a	August 2010	August 2010
Year 3 (2010–2011)	May 26, 2011 (reading only)	August 2011	August 2011
Year 4 (2011–2012)	May 24, 2012 (reading and math only)	August 2012	August 2012
Year 5 (2012–2013)	n/a	May 17, 2013	May 23, 2013

a. Reporting of Results

i. NeSA Reports and Web-Based Reporting

a) Parent/Guardian Report

DRC is committed to developing reports that reflect the needs of NDE and the State of Nebraska. The design of the parent/guardian reports will be **user-friendly, easy to interpret,** and feature **clear graphics** to represent various data elements. We will report overall test results, as well as content area sub-scores at the strand level. Specific reporting information will be determined and approved by NDE.

The hardcopy parent/guardian report will include at a minimum, with NDE's approval: purpose of score report; name of each student; date of test; listing of standards and objectives tested; highest score possible per standard and objective; total number and percent of items answered correctly per standard and objective; performance level achieved for content area; and personalized performance information. DRC plans to include all subject areas in which a student tested on one report.

Our flexible report design can clearly present an array of data elements, including required assessment data, in graphical and narrative formats. Upon contract award, DRC will begin work with NDE to gather requirements for the report design and offer suggestions based on our extensive reporting experience. DRC is excited to assist in creating reports that are psychometrically sound, instructionally sensitive, and meaningful for students, parents, and educators in Nebraska. Please see *Appendix G* for sample student/parent reports. We anticipate that the NeSA parent/guardian report will be similar in nature and design to the Alaska sample we have included.

b) and c) Classroom Roster and Classroom Summary

Classroom Rosters will provide individual student-level results for each classroom per school. The report will include: purpose of score report; name of teacher; date test taken; number of students tested; listing of students in alphabetical order by last name; performance levels for each standard and objective tested; student and indication of other demographic criteria as specified by NDE. We understand that the Classroom Roster may also include item-level results for released items.

Classroom Summary reports will provide teachers with information for assessing overall class performance of all students as a class. The reports will contain: purpose of score report; name of teacher; number of students tested; listing of standards and objectives assessed; median percent of items answered correctly per standard and objective; and number and percent of students at each performance level. We understand that the Classroom Summary reports may also include school, district, and state comparisons.

Classroom Rosters and Classroom Summary reports will only be provided in electronic format through DRC' secure, online Report Delivery System (see *Subheading 8.a.x*).

d)-f) School, District, and State Report Packages

DRC will work with NDE to determine the format for all summary reports. We are committed to providing accurate summary reports. Our overriding goal is to provide useful information to NDE, schools, and districts. We have extensive experience providing accurate, user-friendly, clear, and easy-to-interpret and use reports. Results will be clearly tied to Nebraska standards.

We will provide aggregated and disaggregated data at the school, district, and state levels. Each report will clearly identify the intent of the report, the information included, and which student population(s) is represented. For each school and district and for NDE, we will provide a straight-forward, useful district and state comparison of results. DRC will work with NDE to determine subpopulation categories for disaggregation purposes. DRC has extensive experience providing similar reports; we routinely provide reports that comply with federal and client states' reporting regulations.

The School Report Package will contain whole school achievement level results, NCLB-required subgroup results, and subscore results, as specified in the Table of Test Specifications provided by NDE. DRC also understands that the School Report Package may also include selected results from released items; district and state comparisons; and comparisons with previous years via a Web-based report and database.

The District Report Package will contain whole district achievement level results, NCLB-required subgroup results, and subscore results, as specified in the Table of Test Specifications provided by NDE. DRC also understands that the District

Report Package may also include selected results from released items; state comparisons; and comparisons with previous years via a Web-based report and database.

The State Report Package will contain statewide achievement level results, NCLB-required subgroup results, and subscore results, as specified in the Table of Test Specifications provided by NDE. DRC also understands that the State Report Package may also include selected results from the released items, and comparisons with previous years via a Web-based report and database.

All summary reports will be designed by DRC in collaboration with NDE. Summary reports will be provided in electronic format only via DRC' secure, online Report Delivery System, described under *Subheading 8.a.x.*

g) District Confidential Student-Level Database

DRC will also provide each district with a confidential student-level database containing school identifying information; student identifying information; demographic information; raw item responses for released items; questionnaire responses; raw score totals; scaled scores; and performance levels. We are accustomed to providing similar data to our state clients and will ensure that all secure data file exchange procedures are followed when transferring the database to NDE.

ii. Parent/Guardian Reports

Following the printing of the Parent/Guardian reports (2 copies per student), they will be assembled by school and district, placed in boxes, and labeled "Test Results Enclosed—OPEN IMMEDIATELY." The packaged reports will be shipped directly to districts for distribution to schools. The reports will be packaged and clearly labeled so they can be easily distributed by building/class. Detailed procedures for report assembly will be developed by DRC for NDE's approval. Please see above for reporting timelines.

After reports are packaged, a random sampling quality control procedure will be performed again by checking all of the above in addition to:

- Verifying correct packaging (all reports for a district/school are boxed separately and the correct district/school name is on the outside of each box).
- Verifying that correct mailing address labels are affixed to the outside of each box.

The assembled reports will then be sent to the districts by UPS. In addition, DRC's Project Management Team will monitor the delivery schedule of reports. Each district will sign for its shipment. DRC will track each delivery and compile a record of each signed-for shipment. If a shipment is not delivered within the expected window, DRC's Project Management Team will contact UPS and trace the shipment, providing an update and resolution to the district. Please see *Subheading 4.d, Shipping Requirements for Paper/Pencil,* for more information regarding DRC's packaging and shipping processes and procedures.

iii. Secure Access to Web-Based Reports and Data

DRC proposes our secure, Web-based Report Delivery System for the electronic delivery of NeSA classroom, school, district, and state reports and data files. This system provides schools, districts, and NDE the advantage of receiving school, district, and state reports electronically by selecting reports in a PDF format and data files in Excel, fixed text, or CSV format. DRC will work with NDE to define and finalize the reports and files to be posted on the system.

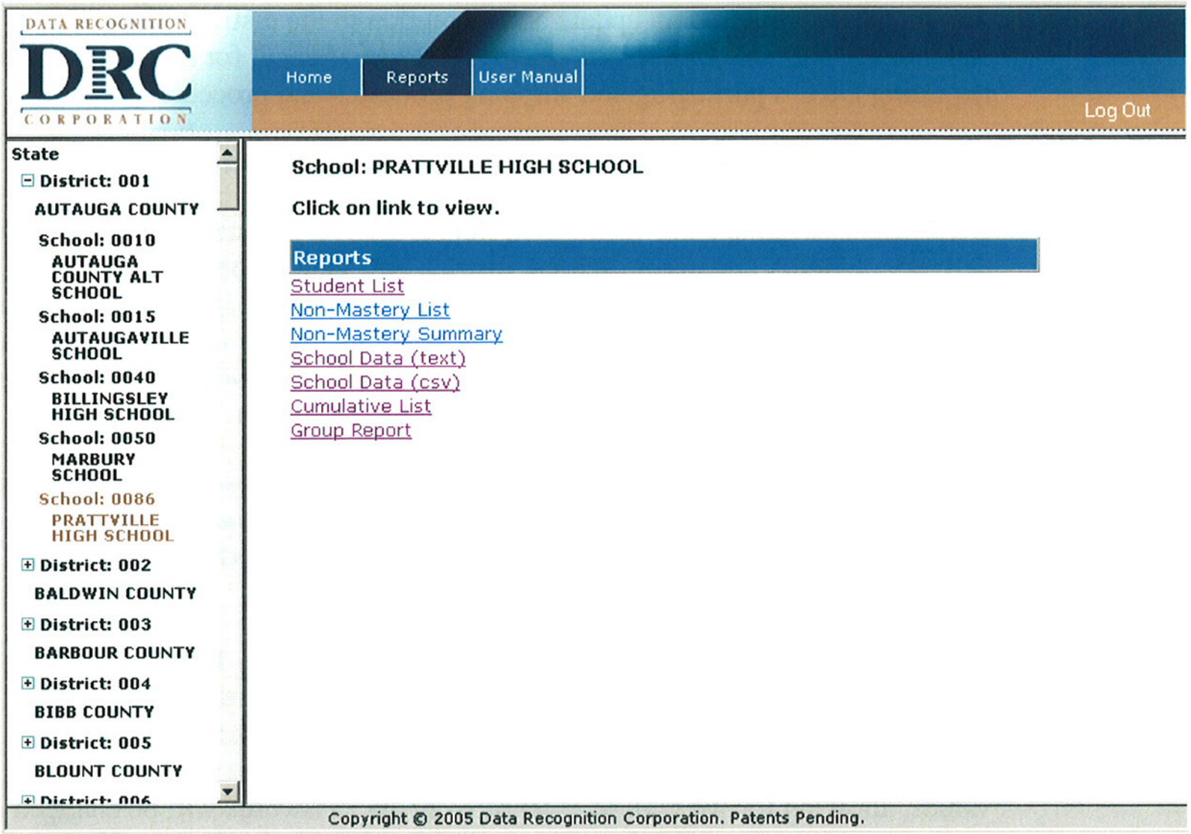
Our Report Delivery System is currently being utilized by many of DRC's clients. As with all of our systems, **the Report Delivery System was designed with ease of use in mind and follows graphical user interface standards, usability guidelines, and security requirements.**

The secure system requires each user to enter a unique user ID and password to ensure confidentiality. User IDs and passwords will be distributed by DRC, according to a process defined in conjunction with NDE. Additionally, all passwords generated will consist of varied case alpha characters and numeric values, allowing for the highest level of security. Passwords will be changed as needed by contacting a DRC representative. During log-in, the user ID and password will be authenticated prior to allowing the user to view reporting results. Note that if NDE chooses to implement DRC's Web portal, eDIRECT, for the NeSA, users would access the Report Delivery System through that portal.

Each user, depending on user ID and password, will have the ability to access different levels of information. The access levels are school, district, or state:

- For a **state**-level user, a list of all districts and the associated schools will be displayed for report/file selection and viewing.
- For a **district**-level user, a list of all schools within a particular district will be displayed for report/file selection and viewing.
- For a **school**-level user, only the school associated with the log-in will be displayed for report/file selection and viewing.

Figure 4–76 displays a screen from this system.



School: PRATTVILLE HIGH SCHOOL

Click on link to view.

Copyright © 2005 Data Recognition Corporation. Patents Pending.

Figure 4-76. Sample Report Delivery System Screen

To ensure the accuracy and reliability of the Report Delivery System, DRC's Software Quality Assurance Analysts, who are experienced with multiple state assessments and web-based systems, will validate that each page, link, and image displays properly. They ensure that the system follows Graphical User Interface (GUI) standards and functions as designed.

DRC follows our standard Project Delivery Quality Control Process and adheres to the 33 Quality Control checkpoints for processing, scoring, and reporting described by the State Collaborative on Assessment and Student Standards (SCASS) on Technical Issues in Large Scale Assessments (TILSA). DRC will ensure that these specific controls are in place and are strictly followed. As added assurance to the NDE, our **Vice President of Quality, Ms. Lisa Peterson-Nelson**, will **conduct regular, formal, documented audits of our quality processes to ensure compliance to procedures**. Any divergence from the requirements will be tracked by our corrective action system and resolved as quickly as possible. All audit results will be utilized as a continuous quality improvement process. TILSA-approved quality checks will be executed to verify that system and school content is correct and report data is 100-percent accurate.

All website changes and modifications will be tested on a dedicated test server before being released into the production environment. The Report Delivery

System website will be tested on various computer platforms, using multiple browsers and numerous browser versions to ensure compatibility with the majority of the general public. Once moved to the production server, the Quality Assurance Analysts will again verify that the Report Delivery System website is accurate and ready for access. DRC will demonstrate the Report Delivery System website to NDE prior to installation.

iv. Archival Copies of Web-Based Reports

For archival purposes, DRC will provide NDE with a hardcopy of each web-based report provided to schools, district, and NDE.

v. Aggregated Results Data File

DRC will provide NDE with an electronic data file that contains all aggregated school, district, and state results provided in the school, district, and NDE web-based reports.

Our Software Quality Assurance staff will ensure the quality of school, district, and state data and make certain that each record is verified for completeness and accuracy. Quality checks will be performed on the data placement and data file formatting for each data element to be displayed on the reports. All data elements will be verified back to the production data file and the data processing rules.

Senior Software Quality Assurance Analysts will conduct a second review to ensure methodology, processes, and procedures are followed and verify that the data files are approved prior to report production.

Data Validation

DRC's experts will conduct quality checks on all data. Quality control checks (please see Figure 4–77) will be performed throughout the system-level testing, including checks of imported and reported data results, to ensure the integrity of the data.

- ***Duplicates***—All systems will be checked for duplicate records and items.
- ***Data standards***—Standard database and data naming conventions will be established and used.
- ***Database accuracy***—Quality assurance staff will perform extensive tests to ensure all data is stored in a secure database environment.

Figure 4–77. Data Quality Procedures

Data File Development

All data file development will be done in close association with NDE to ensure reporting requirements are met. Each data file and report will be **quality checked** for accuracy and completeness a **minimum of three times** by Software Quality

Assurance Analysts and Project Management staff prior to completion. We recognize the importance of this function and have embedded quality checks throughout. Our standard procedures are outlined in Figure 4–78. DRC will work with NDE to confirm these procedures and will modify the process as appropriate.

- ***Record Count Check***—Confirm expected record count.
- ***File Count Check***—Confirm the number of files.
- ***Duplicate File Check***—Verify that duplicate files were not created.
- ***Date/Time Stamp Check***—Verify that the files match the expected date/time stamp.
- ***File Type Verification Check***—Verify that data matches the format specified.
- ***File Log***—A log of files developed will be maintained.
- ***Data Validation***—DRC's Software Quality Assurance staff will use data checking procedures to verify the data is in the specified file layout and matches the expected values.

Figure 4–78. Data File Development Quality Control Procedures

Data File Layouts

DRC will work with NDE to determine appropriate file layouts for each administration. Our expertise in understanding assessment data requirements enables us to provide logical, well-organized, and consistent file layouts.

File layouts will include field names, field descriptions, field values, and starting and ending positions. We will follow an established change control process and track all changes that are made to the layouts. DRC will notify NDE of any changes to the file layouts and provide revised layouts for final approval.

Data Exchange Quality Control Procedures

The exchange of data between entities is a critical and essential component in the success of the NeSA program. To support this process, DRC proposes using our standard data exchange procedures to ensure that all data files are successfully and accurately transferred between DRC, NDE, and others.

We recognize the importance of this function and have embedded quality checks throughout. DRC will work with NDE to confirm these procedures and will modify the process as appropriate.

Data File Delivery

DRC will provide NDE with all required data files and accompanying data file layouts, which will be posted to a secure NDE FTP site or through our Report Delivery System. DRC will work closely with NDE to ensure an easy and accurate data file transfer process is established.

vi. Confidential Student-Level Data File

Following each operational administration, DRC will provide NDE with a student-level data file containing all information and data available for Nebraska students eligible for NeSA testing. The file will include the following data elements, at a minimum: NDE Student ID, demographic information, test form, raw item responses, scored item responses, accommodation information, raw score totals, scaled scores, and performance levels. Prior to the first operational administration of the NeSA, DRC will work with NDE to confirm the data elements and file layout for the student data file.

DRC ensures that all student data remains confidential and secure at all times. Mockups and samples will be provided with a nonspecific identifier (e.g., Student 01). This has been the practice at DRC over the years and is documented as part of the requirements and quality assurance testability for each deliverable.

We incorporate rigorous quality assurance activities throughout the process to ensure the highest level of data quality, integrity, and security. All precode and results data will be accurately stored in a secure database environment. In our computing environment, DRC utilizes security controls that relate to our hardware, data, and network. DRC manages multiple terabytes of client data; therefore, security is an inherent, inextricable, and indispensable component of our system. DRC has extensive experience in designing systems for our clients that have built-in audit trails. All NeSA systems will include audit trails. Each NeSA project team member will be keenly aware of this requirement when managing data requests. Please refer to *Subheading 4., Test Security,* for more information on DRC's security features and procedures.

vii. Interpretive Materials

DRC will annually design, produce, and disseminate interpretive materials for students, parents/guardians, and educators to provide basic information on how to read, interpret, and use the reports. Report interpretation guides will help ensure a clear understanding of each student's NeSA performance, as well as the performance results of schools and districts. The report interpretation guides will provide a sample of a report and a narrative explanation. Sample reports and explanatory text will illustrate results for a fictitious student and/or fictitious school or district. In addition to providing information on how to interpret score reports, the guide will include general information about the purposes of the NeSA and ways to help students, schools, and districts wanting to improve their performance.

DRC will ensure that each proof is free of typographical and format errors before submission to NDE for review. The report interpretation guides will be provided as electronic files (in PDF format) for display on NDE's website. Please see *Appendix G* for sample interpretive materials.

viii. Release of Sample Items

As required by the RFP, DRC agrees to develop materials related to the release of sample items after each test administration. These materials will benefit teachers, parents, and students to help them become familiar with the NeSA items types. In addition, they will provide insight as to how the Academic Content Standards are measured and reported. They will provide teachers with a guide for creating their own items or assessments for use in the classroom.

DRC will include test items, item documentation for mapping items to assessment targets, and scoring materials. These materials will be provided to NDE in a web-based format for posting on NDE's website and/or DRC's Web portal, eDIRECT, dependent on NDE approval. We understand that securing permissions for materials (e.g., reading selections) will be the responsibility of NDE. Selected items will be representative of the breadth of the reporting categories and be determined in consultation with NDE.

It is recommended that DRC Test Development specialists select approximately 25 items per grade level which represents the content of each operational form. The items will be submitted for NDE approval. DRC's Lead Psychometrician will work with DRC's Test Development staff to ensure all recent psychometric and empirical findings are considered when selecting items. Released items will be provided only for the regular assessment.

The released items will not be used on future NeSA test forms and will be marked in the item bank as "used/released." After the item bank is updated, the resulting distribution of items will be reviewed so as to inform, guide, and determine future item development based on the loss of the released items from the bank.

An example of an item sampler with released items that DRC has produced for Pennsylvania is included in *Appendix D*.

ix. Annual Technical Report

DRC will produce an annual Technical Report of the NeSA, three months after the spring assessment results in operational years and three months after the completion of scoring in the standalone field test years. The Technical Report will contain information on the accommodated assessments as well. DRC believes its technical documents represent the best the industry has to offer. The Technical Report will serve as the primary vehicle for documenting reliability and validity evidence for the NeSA and will demonstrate that each of the assessments, and the set of assessments as a whole:

- Serve their intended purposes.
- Are aligned with the test blueprint.
- Fulfill the Table of Test Specifications supplied by NDE (including accessibility criteria).

- Meet or exceed accepted professional standards in educational testing.

DRC will meet with NDE to collaborate on the contents and format for the Technical Report. From the earliest stages of projects, DRC psychometricians are mindful of technical reporting and consider documentation needs continually. The same staff members who plan and conduct project analyses also prepare the associated technical documentation. As with project deliverables, the project's Lead Psychometrician oversees Technical Report preparation. DRC will continue ongoing technical documentation outside of this report as designated by NDE.

Contributions by other functional groups are managed with help of our Project Management team. *Appendix H* contains an example of two DRC's Technical Reports. DRC will deliver the Technical Report in a web-based format for posting to Department websites.

DRC is continually seeking to improve processes. This includes preparation of technical documents. One example of this is the implementation of an internal review of technical documents by independent senior staff members. Cold reads by editors are used to eliminate errors associated with grammar and style.

DRC typically establishes a comprehensive core text for technical reports during the first year of a project. Thought is given to minimizing the amount of new text required yearly and to keeping text that requires modification to established locations. Whenever possible, program output is placed directly into technical documents to limit errors that might occur otherwise. Staff also use visual checks between statistics reported in technical documents and original program output.

As discussed earlier in this proposal, in observance of the demand for quality assurance in the testing industry, DRC employs a **Director of Psychometric Quality, Dr. Adisack Nhouyvanisvong**. To ensure the accuracy and completeness of the NeSA Technical Reports, Dr. Nhouyvanisvong and his team will work alongside the psychometricians and statistical analysts, checking for internal and external consistency and reasonableness. This, in conjunction with the tests and checks performed by our Software Quality Assurance Department, promises Technical Reports that will meet the highest standards. The purpose of the report is to document the entire assessment process in sufficient detail to assure the NDE that the needs of the state educational system are being served and to allow external evaluators to assess the overall quality of the program. It should, for example, be an important document in meeting Federal requirements.

DRC anticipates a solid working relationship with NDE and the Nebraska technical advisory groups.

The following pages include an example of a table of contents for the NeSA Technical Report. DRC will typically provide the assessment's purpose, test blueprint and test maps, Table of Test Specifications, test development procedures, reliability and validity results and graphics, scaling information, inter-rater agreement data, accommodations and testing of students with special needs, security information, administration details, scoring and equating procedures and results, standard setting results, reporting, and appropriate/inappropriate uses and interpretation of data. Appendices will include related materials, administrative regulations, state standards, sample items, committee rating forms, frequency/percentile distributions, state and system performance summaries by ethnic group, and other pertinent information in compliance with NDE. The Technical Report will be delivered in web-based format for posting to the Department websites.

The Nebraska Student Assessment (NeSA)

Sample Table of Contents for Technical Report

Preface: Overview/Purpose

- Assessment Activities in the 2008–2009 School Year

1.0 Background of NeSA Program

- Statewide Testing and Accountability
- Purpose of NeSA Program
- Organizations and groups involved

2.0 Test Development

- Overview of Assessment Test Specifications
    - Test specifications for each subject

3.0 Item Development Process

- Analysis of Bank
- Test Blueprint
    - Test Maps
- Test Development Considerations
    - Item Data Review
        - Differential Item Functioning (DIF)
- Forms Construction

4.0 Test Administration

- Security
- Assessment Accommodations
    - English Language Learners (ELL)
    - Braille
    - Large Print

5.0 Test Administration Procedures

- Test Sessions, Timing, and Layout
- Shipping and Delivery Procedures
- Packaging and Delivery of Materials
- Materials Return
- Test Security Measures
- Assessment Accommodations
- Online Test Administration

6.0 Processing and Scoring
- Receipt of Materials
- Scanning of Materials
- Scoring of Multiple-Choice Items
- Training
- Security

7.0 Scaling, Calibration, and Item Analysis
- Rational
- The Rasch Measurement Model
- Scale Scores and Transformations
- Cut Points for Performance Levels
- Field test analysis

8.0 Equating

9.0 Reports

10.0 Reliability
- Coefficient Alpha
- Internal Consistency
- Standard Errors of Measurement
- Subgroup Reliabilities

11.0 Validity
- Content- and Curricular-Related Evidence
- Construct-Related Evidence
- Criterion-Related Evidence
- Validity Evidence for Different Student Populations
- Secondary analysis studies as determined by NDE

12.0 Standard Setting/ Validation (as applicable)
- Contrasting Groups study
- Results

13.0 Other Studies
- Online vs. Paper Comparability Studies

14.0 Spanish Language Assessments (Option)

15.0 Quality-Control Procedures
- Test Development
- Administration
- Scoring
- Psychometric Services
- Overall

16.0 Glossary of Terms

17.0 References

18.0 Appendices

x. Providing State, Districts, and Schools with Software for Analyzing and Producing Reports

We are pleased to offer Nebraska, as a separate cost option, CAL's interactive Data Analysis Reporting Tool (DART) software to enable the State, each district and school to analyze and produce customized reports from their student-level data. The powerful yet intuitive DART system is designed to allow district personnel to drill-down to specific student subgroups to identify areas of academic weakness. While highly flexible and easy to use, the scalable DART system ensures secure access to sensitive student data as multiple simultaneous users can create instant reports for their desired purpose.

Overview

DART provides a dynamic suite of reporting and analysis tools via a user-friendly interface that allows Nebraska educators to create customized reports by selecting from a menu of parameters including report types, student lists, tests, time periods, and all desired and/or specified subgroup classifications. The intent of DART is to provide reports that are relevant beyond AYP, and our analysis and reporting solution gives educators the tools they need for powerful insight into student performance. Currently implemented for use in the state of Kansas assessment program, DART empowers educators to become data-driven decision makers.

Under this option, CAL will deploy a customized version of DART for Nebraska to provide each district and school with software for analyzing and producing reports from their student-level data. DART for Nebraska offers educators drill-down alternatives for disaggregating student data by multiple demographic variables, navigating from summary level data to a reports for specific subgroups, or drilling down from the summary report to explore an individual student's report in greater detail. DART's available disaggregation variables offer many uses for teachers and other education staff. For example, if a school is preparing for the coming school year, it could be helpful to group the students by the coming year's classes or instructors, and then to present these summaries to the instructors.

System Components

The DART system is comprised of a Variable Selection tool, a ResultsReporter tool, and a Data Analysis tool.

Variable Selection Tool

The DART Variable Selection tool interface is intuitive, interactive, and diagnostic, and provides the authorized user with a flexible system for disaggregating student data. This interface can be customized to NDE's needs and specifications, and will be reviewed in detail with NDE upon contract award. All variable and reporting specifications will be documented by CAL and approved by the NDE prior to final deployment of DART system.

ResultsReporter Tool

CAL's ResultsReporter system is a comprehensive suite of reporting components that includes student, performance level, and longitudinal summary reports. These reports are accessed by teachers and authorized personnel through their own customized system menus. The user-friendly design allows users to access the disaggregated reports that can be viewed on the screen and/or printed to paper.

The student summary report provides comprehensive summary statistics including the number of students tested, mean scale score, number and percent of students at each performance level, and percent of students at and above proficiency. Fully interactive, authorized users can select among content areas, assessment years, and demographic variables. DART's drill-down options offer users the ability to disaggregate student populations by multiple subgroup variables, navigate from summary level data to a roster report for a selected subgroup, or drill-down to a detailed report on an individual student's performance.

The performance level summary graphs include histograms and pie charts for the data corresponding to the selections of the user. The percent of students at each performance level, including the percent of students above and below proficiency, are displayed in these graphs that correspond to the content, subgroup variables, grades, and assessment years selected by the user. In addition to summary reports for corresponding groups of students, users can also drill down to individual students. DART's longitudinal summary reports offer users to ability to track student performance across time at both the group and the individual student level. A sample DART report can be found in *Appendix G.*

Data Analysis Tool

In addition to the dynamic and interactive drill-down disaggregated reports generated by DART, the system can be customized for NDE to offer a full set of data analysis features, allowing users to engage in comprehensive data study by converting values such as N counts and raw scores shown on the reports into percentages, summarize by column headings, compare student performance on different score variables, obtain frequency distributions on different variables, and obtain two-way tabulations for selected variables. CAL will work closely with NDE upon contract award to develop these interactive analysis, display, and data export features.

Security

The DART system is powered by the same CAL online assessment system that is the most secure, robust, and reliable test delivery solution available. The DART system, in five years of operational deployment in Kansas, has never had a breach in student data. DART is fully customizable to allow for the restriction of data elements to various users. DART is protected at the user level by passwords that dictate the level of access granted to each particular user. These configurations can be changed at any time by users with administrative access; changes occur in real-time.

xi. Spanish Language Versions of Parent/Guardian Report Templates and Interpretive Materials

As requested in the RFP, DRC has provided separate costs for developing and delivering optional Spanish language versions of Parent/Guardian report templates, as well as a Spanish translated version of the Report Interpretation Guide. TRI-LIN will provide translation services for these optional items. TRI-LIN has been serving the educational community since 1997. They have experience providing precise translations of educational literature and instructional materials (English to Spanish and Spanish to English); transadaptations of standardized tests and other assessment program tools (from English to Spanish); and custom development of passages and items in the Spanish language for standards-based tests that meet the requirements of the *No Child Left Behind Act* (including reading, language arts/writing, mathematics, and science). TRI-LIN will ensure that all translations for the NeSA communicate as accurately and effectively as the English-language original.

b. Reporting Support

i. Reporting Workshops

DRC also will assist NDE staff in training individuals in interpreting scores and utilizing results to impact classroom instruction, using annotated materials from unique items that are unrelated to the operational assessment. DRC will provide consultation to NDE regarding critical issues and effective tactics for the reporting workshops.

We understand that NDE requires 10 half-day reporting workshops to be conducted following each operational administration. DRC proposes that five reporting workshops be conducted in person at geographically dispersed locations within the State of Nebraska and that the remaining five reporting workshops be held via Webcast. Webcast training would save cost and allow flexible scheduling for the convenience of school and district personnel and NDE staff. NDE staff could log into the reporting workshops from their offices and school and district personnel could log in from home or office locations, while DRC staff facilitates the workshops from our corporate office. Using Webcast would allow NDE staff to attend as many reporting workshops as they desire without taking full days away from other responsibilities. The use of Webcast would also reduce the amount of time required of district and school test coordinators to attend the reporting workshops. These reporting workshops would occur no later than one month after the administration of the NeSA assessments. DRC will work with NDE to determine dates and locations for the reporting workshops.

Announcements and registration forms for the workshops will be mailed to schools and districts at least six weeks prior to the workshop dates. A registration form will also be posted on the DRC-hosted NeSA online portal. A database of registered attendees will be developed which will be used to produce sign-in sheets at the workshop, nametags, and, if needed, the ability to print required certification of completion.

DRC will assist NDE in the creation of materials and PowerPoint presentations for the reporting workshops. All workshop participants will receive information about the actual test, as well as useful information to take back to their districts. These materials will be distributed to all meeting participants. For Webcast participants, materials will be posted online for downloading and/or printing. This will be useful for school and district staff unable to attend a reporting workshop.

DRC will make all arrangements for the reporting workshops. DRC's Program Management Team will handle all of the details related to reserving meeting space, providing necessary audio-visual and computer equipment, arranging for the Webcast workshops, and all administrative activities, including notification of participants, material preparation, and on-site and online registration. For the on-site workshops, DRC will be responsible for lunches, refreshments, meeting facility costs, and other costs associated with attendance by DRC NeSA personnel.

DRC facilitators will include DRC project team members experienced with all aspects of report interpretation. During the on-site reporting workshops, DRC staff will be available to answer questions from participants. Webcast reporting workshops will include a live chat feature to give participants the ability to interact with each other, NDE staff, and DRC personnel. We have prepared reporting workshops for many other assessment clients and look forward to working with NDE staff to develop similarly successful reporting workshops for the NeSA.

ii. Toll-Free Customer Support

DRC will provide customer service support of experienced, informed, and responsive professionals who understand all aspects of the NeSA program. Our customer service function is organized such that only **staff trained in the NeSA program, including reporting, will respond to calls**. As part of the training process, a program-specific customer service manual will be developed that includes frequently asked questions and responses, a program overview, and information on due dates, etc. This manual will play a pivotal role in standardizing the customer communication process for this program. The use of this manual will ensure that Nebraska test coordinators and personnel who call DRC are provided with accurate and consistent information.

DRC's highly experienced and trained customer service staff will be **available throughout the entire contract period** to answer calls on the toll-free number from 8 A.M. to 4 P.M. CST each day. To enhance our service during peak times, for the weeks surrounding the release of assessment results, including at least three weeks following result distribution, the toll-free number will be staffed from 7 A.M. to 5 P.M. CST. Almost all issues will be resolved within 24 hours; callers with complex situations requiring additional time for resolution will receive regular updates on the status of their issues until resolution is complete.

In the unlikely event that telephone service is interrupted, DRC will send an email notifying assessment coordinators that the telephones are down and will send another email once service has been restored. In addition, DRC customer service representatives will have access to cell phones that can be used in emergency situations. In extremely rare cases, if no representatives are available, callers will be able to leave a voicemail message. Test coordinators and NDE staff will also have the option of contacting DRC customer service staff through email and fax. All messages will be responded to within one hour, unless during a NeSA test window (and the weeks before and after), when messages will be returned within 30 minutes.

Please see *Subheading 4.e.ii., Toll-Free Customer Support,* for a thorough discussion of our customer service function.

iii. Investigation of Reporting Errors and Discrepancies

All communications regarding possible reporting discrepancies and errors will be captured and maintained in DRC's customer service database (please see *Subheading 4.e.ii.* for a thorough discussion of our customer service function and customer service database system, EPIC). All instances of possible reporting discrepancies and errors will be immediately reported to NDE.

If the discrepancy or error involves individual students and a request for rescoring or reprocessing, those requests will be submitted to NDE for approval prior to rescore processing. After approval is received, DRC staff will initiate and track the retrieval and rescore process. All rescores will be scored manually by experienced and qualified personnel. All applicable security and quality-control procedures that were implemented by DRC during the original processing and scoring will be maintained (please see *Subheading 4., Test Security,* for more information on DRC's test and data security procedures and *Subheading 5., Scanning/Imaging, DRC's Quality Management System,* for more information on DRC's quality assurance procedures). Rescores will be completed within 10 business days of receipt of NDE rescore approval. Reprocessing and rescoring will be available for 120 days after schools and districts receive their test results. DRC reserves the right to charge for rescore requests, except in the event that any materials have been inaccurately processed, scored, or reported, in which case DRC will retrieve and reprocess them and provide replacement reports and data files at our own cost. Upon contract award, DRC will work with NDE to determine a mutually acceptable charge for rescore requests.

DRC's quality management process focuses on issue prevention to ensure that processing and reporting errors are not made. If a processing error is discovered, DRC will perform all analyses necessary to correct the error prior to reporting of results. If an error in scoring, analyses, or report development is discovered, DRC will notify NDE immediately and provide a solution to remedy the error.

c. Retrieving Student Work

All retrieval requests will be submitted to the NDE project manager for approval prior to processing. After approval is received, DRC staff will initiate and track the retrieval process. Student answer sheet images, original, processed answer sheets, printouts of results, and/or reports can be retrieved quickly and efficiently as the need arises, either during or upon completion of processing. Hardcopies of these materials will be easily retrievable because of DRC's effective document storage procedures (please see below). Additionally, DRC's IBML image scanners allow for on-demand retrieval of specified images (e.g., specific batch files, specific grades, specific students); each image is assigned a unique identification number that allows for quick and easy retrieval at the student and school level.

Depending on NDE preference, either paper or electronic copies of these materials an be provided to the requesting party. Electronic copies would be available as PDF files on CDs or other desired media; the PDFs would be viewable using Adobe Acrobat Reader.

Secure Material Storage

Upon completion of processing, answer sheets and other secure test materials are securely boxed and stored at DRC's secure processing facility for a period of 120 days after schools and districts receive their test results.

Processed answer sheets and other testing-related materials can be retrieved quickly and efficiently as the need arises, either during or upon completion of processing. Materials will be retrieved within two business days of DRC's receipt of official NDE, school, or parental requests. The following steps will ensure the quick retrieval of documents:

- Project-specific box labels will be created containing the following information, as applicable: unique customer and project information, materials type, batch number, pallet/box number, and the number of boxes for a given batch.
- Boxes will be stacked on project-specific pallets. Each pallet will be labeled with a list of all the batches it contains.
- Before each pallet is stored, a quality check will be performed to ensure accurate boxing and pallet content labeling.

After the initial 120-day period, the secure test materials will be moved to an off-site secure facility. This facility will be climate- and pest-controlled, allowing for the preservation of the documents. The documents will still be able to be retrieved, using the above organization scheme. The documents will be retained for at least one year following the last date of the assessment window.

Electronic images and data will be stored for three years following the receipt of test results by schools and districts. The storage system will allow efficient and easy retrieval of individual student tests, data, and results/reports within a short timeframe. Materials will remain secure until written authorization has been received from the appropriate NDE contact to release or securely destroy secure test materials, including answer sheets or images.

Cost and Timetable for Retrieving and Delivering Student Test Documents

DRC will retrieve and deliver images of student answer sheets from the current administration within two business days of receiving an official request. For past administrations, DRC can retrieve and deliver images within five to eight business days. DRC will retrieve images of individual student answer sheets for the most recent administration at no cost. DRC will charge a fee for retrieving images for administrations older than one year, to be determined in conjunction with NDE upon contract award.

9. STANDARD SETTING

a. Standard Setting

In this section, DRC presents the proposed plan to complete both the standard settings and standards validation of all subjects included in the NeSA assessments. Student level results will be reported indicating an overall level of performance according to the three achievement levels established by NDE. All standard settings will be conducted the summer following the first operational administration of each of the NeSA assessments. The table below, Table 4–11, details the schedule for these events.

Table 4–11. Standard Setting Schedule

Subject	Standard Setting Year	Standards Validation Years
Reading	2010	2010 and 2011
Mathematics	2011	2011 and 2012
Science	2012	2012 and 2013

DRC believes the standard settings for reading and mathematics should take a week and science should last three days. DRC plans to run the standard setting with three grade groupings (elementary, middle, and high school) concurrently starting with grades 5, 6, and 11 and ending with grades 3 and 8. Grade 11 in reading and mathematics would be completed in a shorter time frame, ending after three days. It is important that all panelists be trained together to maintain consistency and coherence.

i–ii. Standard Setting Methodology and Procedures

DRC will use the Bookmark standard setting method to set standards for all subjects of the NeSA. The Bookmark procedure (Lewis, Mitzel, & Green, 1996) is appropriate for this project, as items can be reliably ordered by difficulty. In addition, the task required of the judges is considered less complex than the tasks required by other methods. Judges are asked to determine cut score(s) based on this difficulty scale and provide their judgments of items and the separation of one ability level from another.

DRC has successfully conducted standard setting meetings using the Bookmark method for several of state clients (e.g., Alaska, Idaho, Louisiana, North Carolina, and Pennsylvania). A one-page summary of the process is included in Figure 4–79.

Figure 4–79. Summary of Bookmark Method

Materials

1. **Performance Level Descriptors (PLD):** the best explanation of what it means for a student to be *Proficient, or Basic,* or *Advanced.* This is what you will tell parents when they ask what it all means.
2. **Operational Test Forms**: the tests the students took.
3. **Ordered Item Booklets (OIB):** the same items arranged in order of difficulty, easiest to hardest.
4. **Item Map:** a list of the items in the same order as the OIB where you can make notes about the knowledge, skills, and abilities required by the items
5. **Rating Sheet:** the form on which you record your bookmark placement and that you turn into the staff.
6. **Border-line Student Descriptions:** a description, developed by the group, of what the students who are between levels are able to do. These are students that you would be uncertain about how to place and they will be the focus of your deliberations about where to place the borders.

Process

1. Take the test as though you were a student. Don't work too hard on the constructed response.
2. Discuss in the large group (room) what the **constructed response *(CR)*** tasks require and what the rubrics expect. Use the problem as presented to the student, the scoring rubrics and the exemplar papers for each score point. What additional work is needed to move to the next score point? Make notes of your analysis on your Item Map.
3. Discuss in the small groups (table) what *knowledge, skills, and abilities* are involved in each **multiple choice** *(MC)* item. Use the OIB and work through in difficulty order. What makes each item more difficult than the one before it? Makes notes on your Item Map or OIB.
4. **Round 1**: Placing Your *Performance Level Bookmarks*:
   - Place your first bookmark when you come to the item that fewer than 67 out of 100 borderline *Proficient* students will be able to answer correctly.[1]
   - Place your second bookmark when you come to the item that fewer than 67 out of 100 borderline *Advanced* students will be able to answer correctly.
   - Place your third bookmark when you come to the item that fewer than 67 out of 100 borderline *Basic* students will be able to answer correctly.
5. **Round 2**: Discuss the placement of bookmarks with others at your table. Focus only on the items between the lowest and highest flags at your table. Move any or all of your bookmarks if and only if you think it is appropriate.
6. **Round 3**: Discuss the percent of students expected to fall in each performance level (impacts) in the large group and the placement of bookmarks with others at your table. Focus only on the items between the lowest and highest flags at your table. Move any or all of your bookmarks if and only if you think it is appropriate.
7. Complete the **survey** of your assessment of process and result.
8. Turn in your **materials** and take nothing with you except pleasant memories.

[1] For CR, ask yourself if 67 out of 100 borderline students could do this well or better?

Place the bookmark in front of the first item for which the answer is no.

Methodology

The Bookmark standard setting method has two components: the ordered item booklet (OIB), which presents test items in order of their scale (difficulty) locations as determined by the measurement model calibrations, and the item map, which contains both content and statistical information and is used by panelists to record their individual judgments.

In the OIB, the scale locations (difficulty) correspond, in terms of rank order, to classical item difficulties (*p*-values). The easiest item, based on scale score location, is placed in the front of the booklet, while the most difficult is placed at the back. This approach capitalizes on the desirable features of the scaling techniques, which place both items and students on the same scale; given the assumptions of the measurement model, a student's test performance (i.e., score) provides a theoretically known probability of answering a given item correctly.

A primary feature of the Bookmark standard setting methodology is that panelists can make cut score judgments directly onto the score scale, in the context of item content and grade-level expectations. The panelists place a bookmark in the OIB at the point that divides the item content that a student at a given performance level should know and be able to answer from the item content that is too difficult. In this way, content difficulty is directly related to expectations for student performance.

Following several rounds of consideration, final cut scores are established by determining the median of the cut scores by table, which are computed as the medians of individual panelists within each table. (Medians are generally preferred to means because they reduce the influence of extreme judgments, should any exist.)

As applicable, ancillary materials will be placed under a separate cover in order to facilitate the review of those materials. In addition to the OIB, participants will be provided with an item map and supplies, such as paper and adhesive notes. The item map is a table in which each row represents an item in the OIB, ordered in the same manner, with additional information as follows: (1) the scale location for the item, (2) the content categorization, (3) the source of the item (e.g., form and item number), and (4) space for panelists to record notes.

Project Description

This standard setting will include:

- Setting academic achievement standards and reviewing and validating the achievement level descriptors for the assessments using a valid, legally defensible standard setting plan and method that meets the requirements in the RFP.

- Appropriate training of standard setting committee members in the Bookmark method for purposes of determining standards based on their knowledge, judgment, and use of consequential data.
- DRC staff that will lead and facilitate group discussions, including the processes for the standard setting. This will include the review and revision process as required for the achievement level descriptors for each test and each level ensuring alignment with the tests and their current content standards.
- Statistical tables needed to create impact data and used in the iterative item mapping (Bookmark) standard setting procedure.
- Recommending scale score cut points for each of the content area tests for the three Nebraska achievement levels.
- A Technical Report of the process used to generate the recommended cut points.
- Technical documentation and reporting to NDE on the strategies and procedures used prior, during and after the standard setting. Documentation of standard setting data collected, results of analysis, achievement level descriptors and recommended standards based on committee judgment will be included.
- A brief but timely executive summary containing the recommended cut scores from the panel group, along with the impact data provided to the group soon after the final sessions as noted in this project timeline for this RFP.
- Meeting with NDE staff as required to fulfill the terms of this RFP.

DRC's Proposed Team

Assisting **Dr. Ronald Mead, Lead Psychometrician,** in all subject area standard settings will be **Mr. David Chayer, Vice President of Psychometric Services**. Mr. Chayer will lead the standard setting meetings and has extensive experience in Bookmark standard setting techniques. He has managed and provided training and facilitation for over a dozen large-scale meetings in the Bookmark method. These meetings have included projects for Alaska, Idaho, Minnesota, North Carolina, and Pennsylvania. Mr. Chayer has performed and directed research, psychometric, and test development activities in norm-referenced, large-scale assessment and licensure/certification testing programs for both paper-and-pencil and computer-based testing. He joined DRC in 1999.

Standard Setting Panel

An important aspect of the standard setting procedure is the selection of educators with expertise and experience in the subject area and the relevant grade level students. DRC will work with NDE to ensure that appropriate panelists are recruited and will assist in the recruitment of Nebraska educators for this process. DRC will contact, assemble, and train the members for participation in this process.

Each standard setting committee will be composed of a diverse group of 15 subject area teachers, special education teachers, English language (ESL) specialists, and curriculum specialists in Nebraska who have reviewed items in the past or have been recommended for the standard setting process. DRC acknowledges this group must be familiar with the subject matter (content), the population, the instructional environment, and other variables as determined by NDE. DRC also acknowledges that it needs to select members for the panel who are diverse in gender, ethnicity, and regional residence reflecting the diversity in Nebraska.

The proposed plan will use three separate panels: elementary (grades 3–5), middle school (grades 6–8), and high school (grade 11). Using the same panel for three consecutive grades will help ensure coherent recommendations.

When testing in consecutive grades, it is crucial that the performance standards be coherent across grades. Although this was not initially included in the development of the Bookmarking procedure, it is important consideration in the procedure DRC is proposing. The process begins with grades 5 and 6. When these standards have been tentatively established by separate panels, DRC is proposing to bring the two panels together to discuss the work jointly. The final recommendations will then be developed with the input from the other panel. As the recommendations are developed for the remaining grades (grades 3 and 4 for the elementary panel and grades 7 and 8 for the middle school panel), the panels will be reminded of the joint results of the five-six panels to maintain a consistent pattern across grades.

Materials

The materials that are central to the process include:

- The preliminary achievement (performance) level descriptors, to define what students at each level should know and be able to do, provided by NDE.
- An operational form of the test, to demonstrate how the students experience the assessment. While states vary in whether they provide participants with actual operational test booklets, DRC has found that it is useful for participants to see the items in exactly the same form as students saw them so that participants can experience the test in the same way that it is experienced by the students. DRC feels that the use of operational test

booklets adds face validity to the standard setting process and allows the panelists to feel that their work is set within a real-world context.

- The Ordered Item Booklet (*OIB*), to be used for placing the bookmarks.

The OIB will contain items from the common metric item pool arranged in *location* order. Each multiple-choice item will appear once in the booklet. For any item, all preceding items should be easier and all following items should be harder. The locations will be defined for multiple-choice items by placing each item at the level where a student will have a 67% probability of answering the item correctly.

Bookmark Training

An important aspect of the project will be the participants' understanding of the procedure. One important aspect of the training is the emphasis on the role of panelists to not make judgments about the wording or the difficulty of items. Rather, the role of the panelists is to carefully weigh the knowledge and skill levels necessary to have a 0.67 chance of correctly answering the questions.

Each panelist will receive extensive training in a large-group setting prior to making any recommendations. Panelists will receive an orientation to the Bookmark method and practice the mechanics of the process using a short "practice test" composed of non-secure training materials taken from a public source (e.g., released NAEP items).

The Bookmark Placement Task

Participants express their judgments of cut scores by simply placing a tab or bookmark between the ordered items judged to represent the cut point. A separate bookmark is placed for each achievement level. Training will emphasize the following points:

- The bookmark represents a judgment of the demarcation between items that a student at the threshold of a performance level (the student minimally qualified to attain a given achievement level) should know and be able to do and those the student is unlikely to know or be able to do.
- Bookmark placement should not be thought of as separating two items, but rather two groups of items. In other words, a placement should not hinge on distinctions drawn for adjacent items with similar locations. Rather, the collective locations of the group of items below the bookmark should be compared with the collective location of the group of items above the bookmark.
- Students with a scale score at a given cut score will have approximately a 0.67 probability of correctly responding to a multiple-choice item also at the cut score. These same students will have a higher probability of success on easier items (before the bookmark placement) and a lower probability of success on harder items (after the bookmark placement).

Use of Impact Data

Impact data from the spring operational assessment will be presented to the panelists for the third round of deliberations. These data will consist of the frequency distributions of the students' scores. In particular, it will provide estimates of the number and percentage of students who fall into each of the three performance levels. Although the Bookmark procedure is an item-based method, it is generally useful to provide the impact data to help ground the panelists in the reality of student scores and typically leads to more defensible levels for the performance standards

Bookmark Process

The standard setting process will involve three or more rounds of placing and reviewing the bookmarks. There is no intent to reach a consensus; the panelists will be instructed to place their bookmarks where they believe they should be, not where others in the group believe they should be. The first round will focus on each individual's placement of the bookmarks before discussion. DRC believes that this round will provide the best estimate of the true inter-rater variation.

Subsequent rounds will offer the opportunity to revise the individual bookmarks after increasing levels of feedback. The feedback after Round 1 will include only the locations of the bookmarks for all panelists for each level. This will give the panelists the opportunity to see how their decisions compare to the other members of the group and to discuss the differences. Frequently, differences are traced to differing interpretations of the achievement level descriptors.

DRC plans to work with the NDE staff present at the standard setting to review the results of rounds prior to information being presented to the panelists.

Round 1

The first round of the Bookmark process begins with a review of the ordered item booklets as part of a small group. Participants review each item, ordered in terms of difficulty, and are asked to determine and discuss what subject area knowledge, skills, and competencies are required to correctly respond to each item. In this way, items are directly compared, one to another, in terms of the content and skills that must be mastered for each successively more difficult item.

At this stage, participants are encouraged only to identify those skills that a given item requires for mastery of the underlying content. The Round 1 bookmark placements are made individually and discussion among group members is discouraged. This is intended to ensure that the Round 1 judgments are independent and to try to reduce the influence of others' opinions or the opinion of a dominant group member.

At the completion of Round 1, initial cut scores defining the boundaries between each of the performance levels will be computed by DRC staff.

Round 2

Panelists will begin Round 2 with an extensive discussion of their Round 1 ratings. This discussion typically begins at the small group level, led by the table leader. The discussion centers on what students should know at each of the achievement levels. Results of the Round 1 judgments will be presented to the panelists at the beginning of Round 2, including a list of the Round 1 bookmark placements made by each panelist at the each of the tables.

The results from the Round 1 judgments form the basis for the initial discussion phase of Round 2. Panelists will discuss in the small group where they believe the cuts should fall. Following small group discussion, a large group discussion (i.e., across tables) will be facilitated to incorporate more perspectives into Round 1 placements.

After the large group discussion, individual panelists will again review their original bookmark placements and make new bookmark placements. The judgments are entered into a spreadsheet program and the median cut score is calculated for each small group and for the full panel. The latter is used to estimate impact data.

All individual recommendations will then be collected, recorded, and analyzed while the group takes a brief break. Feedback on the overall panel recommendation and the projected impact will be provided to the group as a whole.

Round 3

Panelists will begin Round 3 with extensive discussion of their Round 2 ratings. As in the previous rounds, the judgments from the prior round form the basis for the initial discussion. Each small group will discuss where they believe the cuts should fall between the achievement levels and why.

Following small group discussion, a large group discussion (i.e., across tables) will be facilitated to incorporate additional perspectives into where the levels should be located. Impact data, in the form of percent of students estimated to fall in each performance level, will be provided to help panelists frame the effects of their judgments.

Following the Round 3 large group discussion, panelists will again review their original bookmark placements (in the OIB) and make final bookmark placements. These judgments are once again entered into a spreadsheet program and the median cut score is calculated for each small group, as well as for the full panel. The latter is used to estimate impact data.

All results for all rounds from the week of meetings will be summarized and recorded in a technical report for submission to NDE. Upon approval, DRC will generate the final scale score cutpoints for each test and achievement level.

Evaluation of Standard Setting

After the standard setting is complete, the panelists will be asked to complete a short questionnaire to evaluate the standard setting process. Results of the questionnaire will be included in the technical documentation.

Standard Validation Contrasting Groups

Prior to the standard setting meeting, DRC proposes to involve all Nebraska teachers in a *contrasting groups* study. There are two purposes for this study. First, it will provide first-hand information from the classroom teacher about each student's expected performance on the assessment. Second, it provides a cost-effective strategy to validate the performance standards in the second operational year for each content area. This will avoid the necessity, and all the associated costs and delays, of reconvening standard setting panels in the second operational year. The data collected in the first operational year will provide a basis for evaluating the second year results and to provide feedback and guidance to the panels during the item mapping process.

Using the contrasting groups method, teachers will be given a pre-coded roster of students in their classes and asked to classify each of the students into one of the performance categories: for example; *Not Proficient, Partially Proficient, Proficient* and *Advanced.* A copy of the Performance Level Descriptors (PLDs), which describe what typical students in each of the four performance categories should know and be able to do, will be included with the questionnaire, along with the pertinent question, "Which of these categories best describes each of your students?"

DRC is proposing that teachers be surveyed via an online questionnaire that will be available through the entire testing window and one week after.

Participants

In 2010, DRC recommends that all reading teachers in grades 3 through 8 and 11 be included in the survey. This will ensure an adequate sample and enable DRC to create a representative sample from those who respond allowing for a reasonable level of attrition. Pre-coding should take care of most issues that may arise due to class information linking with teachers.

In 2011, DRC will again survey all reading teachers in grades 3 through 8 and 11. The results of both surveys will then be analyzed and a performance standards validation report for reading will be provided to NDE. In 2011 and 2012, all mathematics teachers in grades 3 though 8 and 11 will be surveyed. In like manner to reading, these two surveys will be analyzed and a validation report for mathematics will be sent to NDE.

In 2012 and 2013, science teachers in the grades selected by the Nebraska State Board of Education (one in elementary, middle, and high school) will be surveyed in a similar manner and a validation study for science will be provided to NDE.

Method/Process

In the online survey, teachers will be asked to classify their students into the three performance-level categories. These classifications will be linked to the students' scores and demographic information. Recommended cut scores will be calculated that provide maximum discrimination among the three levels. The results of these analyses will then be used to determine the percentage of students that would be placed into each of the three levels. Point estimates and standard errors of the placements will be included in the technical report. Information from this study will also be used as a type of impact data to provide guidance to the standard setting panels.

Teachers will receive in the instructions, information on performance level descriptors and information that will aid them in making their choices. All information will be made available in the online survey. It is DRC's intention that final performance level standards that are adopted will be derived from the item mapping process, not the contrasting groups. However, the information obtained from the teachers can inform the standard setting process. In the first operational year, the data can help ground the panelists in the reality of Nebraska education and help validate the process. The comparison between the item mapping result and the contrasting group, in year one, will provide a frame of reference for judging the results of the survey from year two. If the results are comparable across years, it would provide sufficient evidence that the standards from the first year are valid; if the results differ across years, DRC will work with NDE and the National TAC to review the data and discuss whether a re-validation of the item mapping process is warranted.

iii. Support for Standard Setting Activities

DRC will be responsible for all administrative and logistical arrangements and costs for standard setting meetings. This support will include, but not be limited to, maintenance of participant databases; creation of a meeting schedule in collaboration with NDE; mailing of meeting notifications to all standard setting participants; the production of all training, reference and support materials; and facility arrangements including meeting rooms, meals, refreshments, and lodging for participants and NDE staff members. Financial support and travel-related and other relevant expenses for participants and NDE staff members will be provided at the rates agreed to by NDE. Details regarding our assumed specifications for meeting costs related to standard setting can be found in *Appendix J*.

iv. Standard Setting Report

A draft of the technical documentation will be presented to NDE no later than 30 days after the standard setting is complete. At a minimum, this draft will include the following.

- History and Purpose of the Assessment
- Recommended cut scores

- Standard setting method
  - Documentation from NDE on selection of judges
  - Standard setting process
  - Documentation on construction and implementation of materials used during the process
  - Copies of non-secured materials used
  - Training
- Copy of each judge's completed ratings during each phase
- Documentation of feedback received during the process
- The necessary characteristics and concepts of performance at each achievement level
- Copy of each judge's completed evaluation of the process

10. Exit Strategy

At DRC, we have a successful history of cooperation with other providers of large-scale assessments. We achieve this success by placing a testing program's success as a top corporate priority. Through hard work, attention to detail, and a forward-thinking management team, DRC has maintained an excellent reputation in the testing community. The dedication of DRC staff to the ultimate goal of all assessment programs—the improvement of the educational experience of students—ensures that we will find ways to build relationships and solve issues when working with other vendors.

DRC acknowledges that, if not selected to continue work on the NeSA, we will be responsible for transitioning all assessment materials to the new contractor and/or the State upon conclusion of this contract. We understand that our end-of-contract responsibilities would include:

a. Providing a draft detailed Turnover Plan.

b. Modifying the Turnover Plan based on NDE review, prior to contract termination.

c. Transfer data, assessments, reports and other applicable materials in a format prescribed by NDE.

d. Provide technical and professional support to NDE and/or a successor contract in support of the turnover.

e. Prepare and submit initial draft through final deliverables for NDE review, comment and approval.

We will deliver all electronic data files, reports, supporting documentation, and any other materials developed for the NeSA program to the new contractor and/or the State in the formats specified by NDE. We recommend that all transferred data and information will be transferred via CD-ROM/DVD or a secure file transfer protocol (SFTP) site, utilizing all necessary security measures, including encryption. We will also provide staff to work with the new contractor to facilitate the transition of the NeSA program.

CONCLUDING REMARKS

DRC and our assessment partners, CAL and TRI-LIN, are pleased to submit this proposal for the NeSA assessment system. In this Scope of Work, we commit to fulfilling all requirements of the RFP, and have demonstrated the technical procedures that we will adhere to in developing and administering the NeSA. We believe that these procedures, coupled with our highly experienced and dedicated management team, will provide NDE with an unsurpassed level of commitment to a high-quality testing program.

DRC and CAL would also like to extend an invitation to proposal reviewers to visit any of our locations during the review process. It would be our pleasure to demonstrate in person our capabilities and commitment to providing a superior assessment program for NDE and the education stakeholders and students of Nebraska.

WORKS CITED

Allen, N. L., Carlson J. E., & Zelenak, C. A., (1999). *The NAEP 1996 technical report* (NCES 1999 – 452). Washington, DC: National Center for Education Statistics.

Allman, C. (2004). Making Tests Accessible for Students with Visual Impairments: A Guide for Test Publishers, Test Developers, and State Assessment Personnel. (2nd edition.) Louisville, KY: American Printing House for the Blind. Available from http://www.aph.org.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (2nd ed.). Washington, DC: American Educational Research Association.

Andrich, D. (1988). *Rasch models for measurement.* Newberry Park, CA: Sage Publications.

Angoff, W. H. (1971). Scales, norms and equivalent scores, in R.L. Thorndike (Ed.), *Educational measurement*, (2nd edition, pp. 508–600), Washington DC: American Council on Education.

Angoff, W. H. (1974). The development of statistical indices for detecting cheaters. *Journal of the American Statistical Association, 69*, 44–49.

Birch, M. W. (1964). The detection of partial association I: The 2x2 case. *Journal of the Royal Statistical Society*, *B26,* 313–324.

Cook, L. & Eignor, D. (1991). IRT equating methods. *Educational Measurement: Issues and Practice. 10,* 37–45.

Fischer, G. & Molenaar, I. (1995). *Rasch models – foundations, recent developments, and applications.* New York: Springer.

Fletcher, R. B. (2000). *A review of linear programming and its application to the assessment tools for teaching and learning projects.* (Technical Report 5, Project as TTle). Auckland, New Zealand: University of Auckland.

Hanson, B. A., Harris, D. J., & Brennan, R. L. (1987). *A comparison of several statistical methods for examining allegations of copying* (Research Rep. Series No. 87–15). Iowa City, IA: American College Testing Program.

Holland, P. W. & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenzsel procedure. In H. Wainer and H. I. Braun (Eds.), *Text Validity* (pp. 129–145). Hillsdale NH: Lawrence Erlbaum Associates.

Kolen, M. J., & Brennan, R. L. (2004). Test equating, scaling, and linking: *Methods and Practices.* (2nd ed.). New York: Springer-Verlag.

Lewis, D. M., Mitzel, H. C. & Green, D. R. (1996, June). Standard setting: A bookmark approach. A paper presented at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32(2), 179–197.*

Ludlow, L. H. (1986). Graphical analysis of item response theory residuals. *Applied Psychological Measurement*, *10*, 217–229.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719–748.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika 47*, 149–174.

Mead, R. J. (1976). *Assessing the fit of data to the Rasch model through the analysis of residuals*. Unpublished doctoral dissertation, Chicago: University of Chicago.

Mead, R. (1980). Using the Rasch model to identify person-based measurement disturbances. *Proceedings of the 1979 Computer Adaptive Testing Conference*. Minneapolis, MN: University of Minnesota.

Mead, R. J. (2008). *A Rasch primer: the measurement theory of Georg Rasch.* (Psychometrics services research memorandum 2008–001). Maple Grove, MN: Data Recognition Corporation.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danish Institute for Educational Research.

Smith, R. M. (1982). *Detecting measurement disturbance with the Rasch model.* Unpublished dissertation. University of Chicago

Smith, R. M. (1986). Person fit in Rasch model. *Educational and Psychological Measurement, 46*, 359–372.

Smith, R. M. (2000). Fit analysis in latent trait measurement models. *Journal of Applied Measurement*, *1*, 199–218.

Smith, E. V. & Smith, R. M. (2004). *Introduction to Rasch measurement*. Maple Grove, MN: JAM Press.

Smith, E. V. (2004). Evidence for the reliability of measures and validity of measure interpretation, in E. V. Smith & R. M. Smith (Eds.) *Introduction to Rasch measurement*, (pp. 93–122). Maple Grove, MN: JAM Press.

Sotaridona, L. S., & Meijer, R. R. (2002). Statistical properties of the K-index for detecting answer copying. *Journal of Educational Measurement*, *39*, 115–132.

Sotaridona, L. S., & Meijer, R. R. (2003). Two new statistics to detect answer copying. *Journal of Educational Measurement, 40*, 53–69.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal Design Applied to Large Scale Assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Retrieved March 15, 2008 from: http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*, 278–286.

WINSTEPS (2008). WINSTEPS® Rasch Measurement. [Computer Program]. Chicago: WINSTEPS.com.

Wollack, J. A. (1997). A nominal response model approach to detect answer coping. *Applied Psychological Measurement, 21(4)*, 307–320.

Wollack, J. A. (2006). Simultaneous use of multiple answer copying indices to improve detection rates. *Applied Measurement in Education, 19*, 265–288.

Wright, B. D. (1968). Sample-free test calibration and person measurement. In *Proceedings of the 1967 Invitational Conference on Testing Problems*. Princeton, NJ: Educational Testing Service.

Wright, B. (1980). Foreword. In G. Rasch. *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press.

Wright, B. D. & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wright, B. D. & Stone, M. (1979). *Best test design*. Chicago: MESA Press.

DATA RECOGNITION
DRC
CORPORATION